DEVELOPMENT OF MULTIMODAL FUSION-BASED VISUAL DATA ANALYTICS FOR ROBOTIC INSPECTION AND CONDITION ASSESSMENT

by

Tarutal Ghosh Mondal

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Lyles School of Civil Engineering

West Lafayette, Indiana

December 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Mohammad R. Jahanshahi, Chair

Lyles School of Civil Engineering

Dr. Dulcy M. Abraham

Lyles School of Civil Engineering

Dr. Julio Ramirez

Lyles School of Civil Engineering

Dr. Edward J. Delp

School of Electrical and Computer Engineering

Approved by:

Dr. Dulcy M. Abraham

To my parents and family

ACKNOWLEDGMENTS

I have many people to thank for helping me getting through this dissertation. First of all, I express my profound appreciation to my advisor, Professor Mohammad R. Jahanshahi, for his unrelenting patience and encouragement. It has been both a pleasure and a privilege for me to work with him. I would not have made it without his unwavering support, helpful advice, and constructive criticism at every stage of this study.

I wish to thank Professor Dulcy M. Abraham, Professor Julio Ramirez, and Professor Edward J. Delp for serving on my dissertation committee and providing valuable suggestions. I also had the great pleasure of working with Prof Shirley Dyke and Dr. Zheng Wu whose valuable input steered me through this research.

Time has flown by rather quickly. But some of my fondest memories will remain with me forever. I will cherish the pleasurable company of my friends, Sayan Basak, Mohit Singh, Akash Ashapure, Harsh Bohra, Rahul Deshmukh, and Deepak Suthar, among many others, who were indeed like my family away from home. I also had some beautiful times with my academic peers, namely, Rih-Teng Wu, Yu-Ying Huang, Tang Wen, and Zixin Wang. I will carry positive memories of our friendly chats and research collaborations.

I will forever be grateful to my parents, wife, and family for their years of sacrifice and continued support, which have immense contributions to my life. This dissertation stands as a testament to their unfailing love and blessings.

Finally, I would like to extend my sincere thanks to one and all who, directly or indirectly, have lent a hand in this venture.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

CNN	Convolutional Neural Network
RCNN	Region-based Convolutional Neural Network
UAV	Unmanned Aerial Vehicle
RC	Reinforced Concrete
AD	Absolute Depth
RD	Relative Depth
SN	Surface Normal
ADE	Absolute Depth-based Encoding
RDE	Relative Depth-based Encoding
SNE	Surface Normal-based Encoding
RANSAC	Random Sample Consensus
IoU	Intersection-over-Union

ABSTRACT

This dissertation broadly focuses on autonomous condition assessment of civil infrastructures using vision-based methods, which present a plausible alternative to existing manual techniques. A region-based convolutional neural network (Faster R-CNN) is exploited for the detection of various earthquake-induced damages in reinforced concrete buildings. Four different damage categories are considered such as surface crack, spalling, spalling with exposed rebars, and severely buckled rebars. The performance of the model is evaluated on image data collected from buildings damaged under several past earthquakes taking place in different parts of the world. The proposed algorithm can be integrated with inspection drones or mobile robotic platforms for quick assessment of damaged buildings leading to expeditious planning of retrofit operations, minimization of damage cost, and timely restoration of essential services.

Besides, a computer vision-based approach is presented to track the evolution of a damage over time by analysing historical visual inspection data. Once a defect is detected in a recent inspection data set, its spatial correspondences in the data collected during previous rounds of inspection are identified leveraging popular computer vision-based techniques. A single reconstructed view is then generated for each inspection round by synthesizing the candidate corresponding images. The chronology of damage thus established facilitates time-based quantification and lucid visual interpretation. This study is likely to enhance the efficiency structural inspection by introducing the time dimension into the autonomous condition assessment pipeline.

Additionally, this dissertation incorporates depth fusion into a CNN-based semantic segmentation model. A 3D animation and visual effect software is exploited to generate a synthetic database of spatially aligned RGB and depth image pairs representing various damage categories which are commonly observed in reinforced concrete buildings. A number of encoding techniques are explored for representing the depth data. Besides, various schemes for fusion of RGB and depth data are investigated to identify the best fusion strategy. It was observed that depth fusion enhances the performance of deep learning-based damage segmentation algorithms significantly. Furthermore, strategies are proposed to manufacture

depth information from corresponding RGB frame, which eliminates the need of depth sensing at the time of deployment without compromising on segmentation performance. Overall, the scientific research presented in this dissertation will be a stepping stone towards realizing a fully autonomous structural condition assessment pipeline.

1. INTRODUCTION

Buildings form an important part of urban infrastructure systems. Damage in buildings caused by earthquake events not only renders the residents homeless, it brings to a halt various economic activities which are directly or indirectly dependent on building infrastructures. It also disrupts essential service utilities which act as lifelines for the people of the locality. Therefore, it is of vital importance to ensure that the full functionality of buildings is quickly restored in the wake of an earthquake event. The ability of a building to withstand damage and recover in a timely manner following an extreme event is called structural resilience, which is recognized by the scientific community as a promising research area owing to its profound impact on life safety and overall economy. However, an expeditious disaster recovery calls for a rapid and comprehensive evaluation of the nature and extent of damages inflicted by the extreme event on building infrastructure systems. The existing earthquake reconnaissance practices are predominantly manual. A group of certified inspectors visit the affected buildings, taking measurements from the damaged areas, post processing the collected information and finally arriving at the retrofit decision. Needless to say, this procedure is time consuming and expensive as it requires a lot of manpower. Sometimes it also involves risk as the human inspectors need to visit or go very close to a damaged structure which is about to collapse, to record an accurate reading. As a viable alternative, such manual methods can be replaced by inexpensive UAVs or inspection robots which can cruise autonomously through potentially damaged buildings looking for damages and collecting critical information using on-board vision-based and other types of sensor systems which will help identify the problem areas requiring immediate attention. An autonomous engineering assessment like this will identify the risks and mitigate life safety hazards for human inspectors by preventing them from entering a building which is prone to collapse. It will also enable quick evaluation of recovery and repair cost and financial loss induced by downtime. Additionally, reserve capacity for different structural elements can be assessed from corresponding damage levels and retrofit operations can be planned accordingly.

An exhaustive review of existing literature revealed several research gaps, some of which are addressed in this dissertation. It was observed that the previous studies on vision-based

post-disaster reconnaissance of reinforced concrete buildings focused only on surface cracks and spalling, whereas more severe damage categories like spalling with exposed rebars and severely buckled rebars were ignored. This study aimed to fill this information gap by leveraging a deep learning-based approach. A region based convolutional neural network (Faster RCNN) is exploited to detect four different damage types, namely, surface crack, spalling (which includes facade spalling and concrete spalling), severe damage with exposed rebars and severely buckled rebars. The performance of the proposed approach is evaluated on manually annotated image data collected from reinforced concrete buildings damaged under several past earthquakes such as Nepal (2015), Taiwan (2016), Ecuador (2016), Erzincan (1992), Duzce (1999), Bingol (2003), Peru (2007), Wenchuan (2008), and Haiti (2010). Several experiments are presented to illustrate the capabilities, as well as the limitations, of the proposed approach for earthquake reconnaissance. The research outcome is a step forward to facilitate the autonomous condition assessment of buildings where this can be potentially useful for insurance companies, government agencies and property owners.

Besides, civil infrastructures are observed to undergo deterioration over time owing to overloading or unfavorable environmental conditions. This calls for periodic inspection of structures in order to prevent sudden failure or to avoid any untoward human casualties caused by unserviceable infrastructure conditions. The existing inspection techniques are predominantly manual, and consequently time consuming, expensive, subjective, and risky. Numerous studies in recent years focused on autonomous inspection techniques based on the latest advancements made in the areas of computer vision and deep learning [1], [2]. Spencer et al. [3] provides an exhaustive review of available literature on this topic. A number of investigations explored autonomous damage identification from visual data exploiting various image processing [4], machine learning and convolutional neural networks (CNN) [5]–[7] based approaches. Damage quantification also gained some attraction from the research community [8]–[10]. However, most of the previous studies found in literature were invariably agnostic to the time dimension. It is often important to understand how fast a damage is progressing and how long it may take to reach the limit state of collapse. However, disregarding the temporal information in the state-of-the-art damage identification pipeline makes such information scarce, preventing the inspectors act preemptively to

minimize the cost incurred due to the damage. It is therefore necessary to address this knowledge gap existing in this important area of research, which is the focus of this study. A computer vision-based approach is presented for representing time evolution of structural damages leveraging a database of inspection images. Spatially incoherent but temporally sorted archival images captured by robotic cameras are exploited to represent the damage evolution over a long period of time. An access to a sequence of time-stamped inspection data recording the damage growth dynamics is premised to this end. Identification of a structural defect in the most recent inspection data set triggers an exhaustive search into the images collected during the previous inspections looking for correspondences based on spatial proximity. This is followed by a view synthesis from multiple candidate images resulting in a single reconstruction for each inspection round. Cracks on concrete surface are used as a case study to demonstrate the feasibility of this approach. Once the chronology is established, the damage severity is quantified at various levels of time scale documenting its progression through time. The proposed scheme enables the prediction of damage severity at a future point in time providing a scope for preemptive measures against imminent structural failure. On the whole, it is believed that this study will immensely benefit the structural inspectors by introducing the time dimension into the autonomous condition assessment pipeline.

The recent advancements in the areas of computer vision and deep learning have undeniably broadened the scope of vision-based autonomous condition assessment of civil infrastructure. However, a review of available literature suggests that most of the existing vision-based inspection techniques rely only on color information due to the immediate availability of inexpensive high-resolution cameras. However, it should be noted that regular cameras translate a 3D scene to a 2D space which leads to a loss of information vis-à-vis distance and scale. This imposes a barrier to the realization of the full potential of vision-based techniques. In this regard, the structural health monitoring community is yet to benefit from the new opportunities that commercially available low-cost depth sensors like Microsoft Kinect offer. This study aims at filling this knowledge gap by incorporating depth fusion into an encoder-decoder-based semantic segmentation model. A 3D animation and visual effect software is exploited to generate a synthetic database of paired RGB and depth images representing various damage categories that are commonly observed in reinforced concrete

buildings, namely, spalling, spalling with exposed rebars, and severely buckled rebars. A number of encoding techniques are explored for representing the depth data. Additionally, various schemes for the fusion of RGB and depth data are investigated to identify the best fusion strategy. Overall, it was observed that depth fusion enhances the performance of deep learning-based damage segmentation algorithms significantly. In consideration of various practical challenges of robotic depth sensing, this study also investigates different ways to dispense with depth sensing at test time without foregoing the dividend of depth fusion. Moreover, a novel volumetric damage quantification approach is introduced which is robust against perspective distortion.

Scope

This thesis is organized in the following manner. Chapter 2 presents a deep learning-based approach for autonomous post disaster reconnaissance of reinforced concrete buildings. A novel computer vision-based technique is proposed in Chapter 3 to assess the time evolution of a structural defect exploiting historical inspection data. The study presented in Chapter 4 explores the utility of depth information in enhancing the performance of traditional deep learning-based damage segmentation algorithms. Finally, Chapter 5 summarizes the key findings and outlines the course for future research.

2. DDEEP LEARNING-BASED MULTI-CLASS DAMAGE DETECTION FOR AUTONOMOUS POST-DISASTER RECONNAISSANCE

2.1 Background

[1]Buildings form an important part of urban infrastructure systems. Damage in buildings caused by earthquake events not only renders the residents homeless, it brings to a halt various economic activities which are directly or indirectly dependent on building infrastructures. It also disrupts essential service utilities which act as lifelines for the people of the locality. Therefore, it is of vital importance to ensure that the full functionality of buildings is quickly restored in the wake of an earthquake event. The ability of a building to withstand damage and recover in a timely manner following an extreme event is called structural resilience, which is recognized by the scientific community as a promising research area owing to its profound impact on life safety and overall economy. However, an expeditious disaster recovery calls for a rapid and comprehensive evaluation of the nature and extent of damages inflicted by the extreme event on building infrastructure systems. The existing earthquake reconnaissance practices are predominantly manual. A group of certified inspectors visit the affected buildings, taking measurements from the damaged areas, post processing the collected information and finally arriving at the retrofit decision. Needless to say, this procedure is time consuming and expensive as it requires a lot of manpower. Sometimes it also involves risk as the human inspectors need to visit or go very close to a damaged structure which is about to collapse, to record an accurate reading. As a viable alternative, such manual methods can be replaced by inexpensive UAVs or inspection robots which can cruise autonomously through potentially damaged buildings looking for damages and collecting critical information using on-board vision-based and other types of sensor systems which will help identifying the problem areas requiring immediate attention. An autonomous engineering assessment of this sort will help identifying the risks and mitigate

[1]The content of this chapter is published as follows: T. G. Mondal and M. R. Jahanshahi, "Deep learning–based multi–class damage detection for autonomous post–disaster reconnaissance," *Structural Control and Health Monitoring*, 27(4), e2507, https://doi.org/10.1002/stc.2507.

life safety hazards for human inspectors by preventing them from entering a building which is prone to collapse. It will also enable quick evaluation of recovery and repair cost and financial loss induced by downtime. Additionally, reserve capacity for different structural elements can be assessed from corresponding damage levels and retrofit operations can be planned accordingly.

Several studies in the past focused on improving the resilience of infrastructure systems by proposing improved techniques for rapid structural inspection and damage assessment capitalizing on the recent advancements made in fields of computer vision and deep learning. Image processing-based techniques (IPTs) are exploited by many researchers to this end. Yamaguchi and Hashimoto [4] proposed a percolation-based image processing technique for fast and efficient concrete crack detection. German et al. [11] exploited entropy-based thresholding in conjunction with template matching and morphological operations for rapid detection and quantification of concrete spalling during post-earthquake safety assessments. Similar studies related to damage detection in other forms of structural systems include identification of cracks in bridges [12], pavement surfaces [13]–[19], and underground pipes and subway tunnels [20], [21]. Several researchers [22]–[25] in the past also focused on machine learning based techniques (MLTs) for automatic vision-based damage detection where the feature vectors are selected manually in these methods.

On the other hand, rapid enhancement in computational capacity in recent times has triggered a resurgence of deep learning based convolutional neural networks (CNNs) garnering significant attention from the research community cutting across all disciplines. On this account, a number of studies in the past [26]–[28] explored the possibility of applying CNN for efficient and autonomous detection of various structural damages. However, very few studies indeed looked into damage types which are relevant to earthquake reconnaissance of reinforced concrete (RC) building systems, which is the focus of the present study. Cha et al. [29] investigated multiple damage categories such as concrete cracks, steel corrosion, bolt corrosion and steel delamination with the help of region based convolutional neural network (RCNN). However, the damage categories considered by the authors were not closely correlated and are not applicable to RC buildings on the whole. Cha et al. [7] exploited sliding window approach for detecting cracks on concrete surfaces. However, it ignored other types

of damages that are commonly observed in RC buildings post-earthquake events. Yeum et al. [30] recently proposed an RCNN based approach for spalling recognition in RC buildings. This study also does not take into account other damage types that may possibly result in when such buildings are subjected to seismic vibrations. Kim et al. [31] capitalized on image binarization and CNN to distinguish crack from crack-like noncrack noise patterns (e.g. dark stains, shades, dust, lumps, holes etc.) on concrete surfaces. Therefore, this study also had limited scope in terms of varieties of damage types considered. Chen and Jahanshahi [32] proposed a CNN-based approach to detect cracks on nuclear reactors. The false detections were discarded using Naive Bayes data fusion by aggregating information from successive frames in inspection videos. However, this study was also exclusively focused on identification of cracks and other types of damages were ignored. Hoskere et al. [33] harnessed pixel-wise classification of images using deep CNN to identify multiple damage classes in civil infrastructure systems. However, only two (concrete crack and concrete spalling) out of six damage categories considered in this study were relevant to RC buildings and the rest corresponded to deterioration in steel structures and asphalt pavements. For instance, some important damage types such as buckling of column rebars caused by severe earthquake vibrations were not considered in this study. Additionally, it involved expensive training data preparation process like pixel-wise labeling of images. It should be noted that ignoring severe damage categories such as exposed and buckled rebars may have adverse safety ramifications, as it may lead to underestimation of the damage severity and falsely encourage the human inspectors to enter a building which is on the verge of collapse, resulting in fatal injuries. This underlines the necessity of including multiple damage categories representing the entire spectrum of severity in the autonomous damage detection pipeline. Recently, Gao and Mosalam [34] presented a CNN based approach for structural damage classification. Although, this study considered a range of damage categories and various classification modalities, namely, component type identification, spalling condition check, damage level estimation, and damage type determination, it did not focus on localizing the damage in the images. This is an important limitation which was recommended as a part of the further works by the authors.

2.1.1 Scope and Contribution

It is important that the application of deep CNN is extended to multiple damage categories which will immensely benefit earthquake reconnaissance and safety evaluations. The present study aims at filling this gap by proposing a Faster RCNN based detection technique taking into account multiple damage types that may be caused in RC buildings when subjected to earthquake ground motion. Four different damage categories are considered in this study which are – surface crack, spalling (which includes façade spalling and concrete spalling), severe damage (i.e., spalling with exposed rebars) and severely buckled exposed rebars. The CNN architectures that were exploited to this end are Inception v2 [35], ResNet-50 [36], ResNet-101 [36] and Inception-ResNet-v2 [37]. The efficiency of the proposed algorithms is evaluated with the help of earthquake reconnaissance data collected after several past earthquakes such as Nepal (2015), Taiwan (2016), Ecuador (2016), Erzincan (1992), Duzce (1999), Bingol (2003), Peru (2007), Wenchuan (2008), and Haiti (2010). It is observed that Inception-ResNet-v2 performed significantly better than other architectures considered in this study producing a mean average precision value of 63.78%. It is believed that this study will help enhance autonomous post-disaster reconnaissance of RC buildings.

The datasets used in this study for training and evaluation of detection algorithms were collected from different countries representing wide variation in local construction practices and design specification. Therefore, the images contained damages in various shapes, sizes and aspect ratios. This poses a challenge of dealing with this scale variation and devising a detection algorithm which is scale agnostic. This research challenge was addressed in this study by modifying the region proposal network (RPN) [38]. Typically, a 3 x 3 sliding window is applied to the feature map generated by the last convolutional layer in the RPN. At each sliding window location, a number of anchor boxes having different scales and aspect ratios are considered as region proposals to account for scale variability of objects. In the default configuration of faster RCNN proposed by Ren et al. [38], a total of 9 anchor boxes were proposed with 3 different scales and 3 different aspect ratios. However, in this study, seven different scales and eight different aspect ratios were used leading up to 56 anchor boxes, which improved detection accuracy significantly.

Research on robot-based autonomous inspection and condition assessment has many facets. A number of researchers in the past focused on developing advanced robotic systems and path planning algorithms with an eye to futuristic inspection operations. Simultaneously, recent advances in the fields of computer vision and deep learning evoked profound interests in vision-based damage diagnosis which is investigated in this study. The scope of this work is limited to visual data analysis for autonomous multi-class seismic damage identification. Hands-on experiment with real physical robot is not considered here and is a part of future work. The chapter has been arranged in the following order. Section 2.2 discusses the Faster RCNN approach for object detection. Various CNN architectures considered in this study are briefly described in Section 2.3. Image dataset used for evaluation of the proposed approach is presented in Section 2.4. The training scheme and other implementation details are summarized in Section 2.5. The detection results are presented and discussed in Section 2.6. Finally, conclusions are summarized in Section 2.7.

2.2 Faster RCNN

The inception of Faster RCNN can be traced back to introduction of Regions with CNN features RCNN by Girshick et al. [39]. In RCNN, around 2000 category-independent region proposals are extracted from the input image using selective search algorithm. Each region proposal is then sent to a CNN to generate a fixed-length feature vector. Finally, category-specific linear SVMs are used to classify each region proposals. At the end, greedy non-maximum suppression is employed to get rid of the redundant detections. However, RCNN is slow during training and testing since features are extracted from each region proposal in each image and written to the disk. Girshick [40] addressed this shortcoming by replacing the multi-stage training pipeline of RCNN with a single-stage algorithm. In this refined approach called Fast RCNN, an input image together with a set of object proposals is input to a series of convolutional and max pooling layers to obtain a feature map. Then, a region of interest (RoI) pooling layer is invoked to extract a fixed-length feature vector from the feature map for each of the object proposals. Each of the feature vectors is then fed into a sequence of fully connected layers which eventually bifurcate into two collateral output layers

constituting a softmax classifier and a bounding box regressor. Fast RCNN is significantly faster than RCNN during training and testing owing to computation and memory sharing across the RoIs from the same image. It also offered higher accuracy compared to RCNN. However, the region proposal computation was a bottleneck, elimination of which was likely to further speed up the testing process. This was materialized by Ren et al. [38], who proposed a Region Proposal Network (RPN) drastically reducing the cost of region proposal computation at the test time.

2.2.1 Region Proposal Network (RPN)

RPN is a fully convolutional network trained to predict object bounds along with objectness scores. The region proposals generated by RPN are used by Fast RCNN for accurate detection. The RPN and the Fast RCNN modules are unified into a single network enabling sharing of convolutional layers (Figure 2.1). An $n \times n$ sliding window is applied to the feature map generated by the last shared convolutional layer mapping it down to a lower dimension (Figure 2.2). At each sliding window location, a set of k anchor boxes having different scales and aspect ratios are considered as region proposals. The lower dimensional feature is fed into two collateral fully connected layers, namely a box-regression layer and a box-classification layer. The box-regression layer has $4k$ outputs denoting the coordinates of k bounding boxes. On the other hand, the box-classification layer produces $2k$ outputs representing the objectness score of each bounding box. For more details about the training scheme and other implementation details, the readers may refer to the original paper by Ren et al. [38].

On the whole, a CNN is used at first to generate a feature map from the input image. In this study, four different network architectures are exploited to this end which are described in the following section. Subsequently, RPN is used to generate regions proposals; following which Fast RCNN module is utilized for classifying the RoIs and refining the bounding box coordinates. In a way, RPN incorporates 'attention' mechanism telling the classifier where to look. More details about the implementation scheme are provided in Section 2.5.

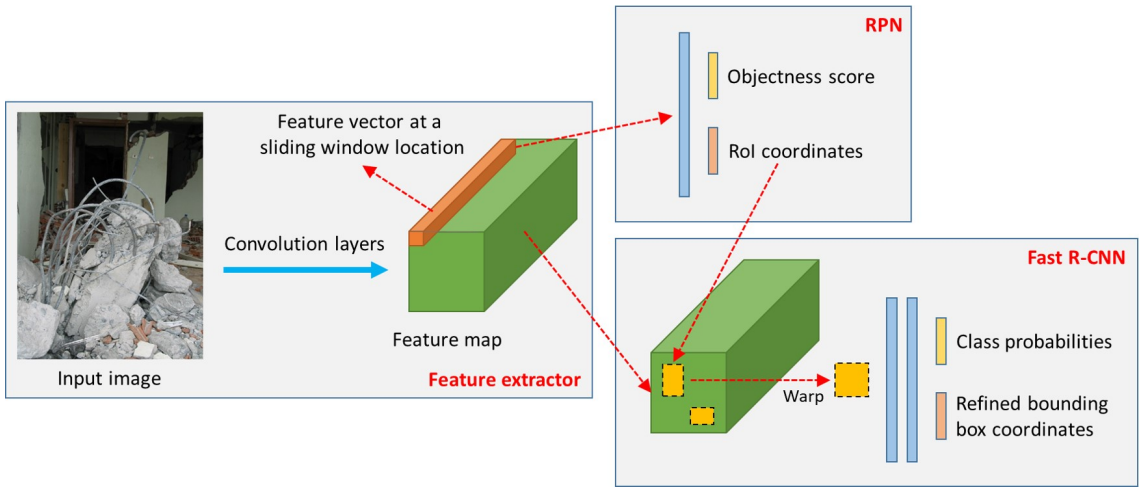


Figure 2.1. Faster RCNN architecture

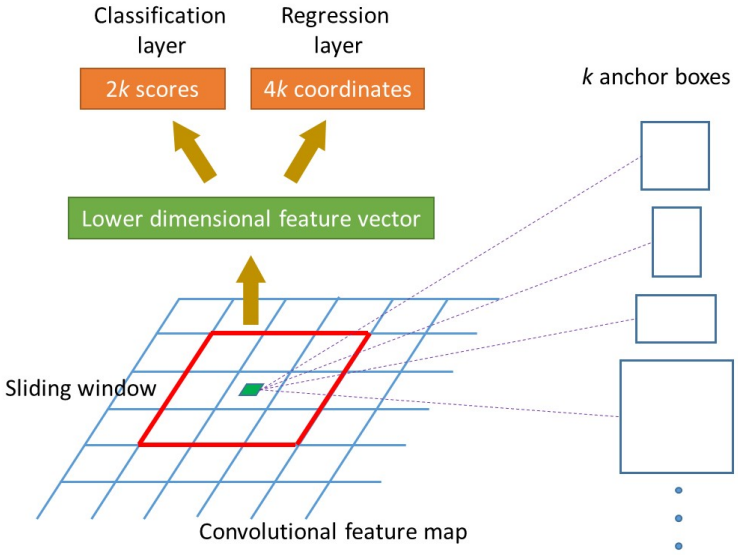


Figure 2.2. Region proposal network

2.3 Network Architectures

2.3.1 Inception v2

Prior to the introduction of Inception network [41], it was a common trend to stack additional convolutional layers to increase the accuracy of the network, leading to a very deep network. Such a deep network is fraught with numerous limitations such as overfitting, vanishing gradients, and expensive computational cost. On the other hand, wide variation in the object size makes it challenging to estimate the most optimum kernel size for convolution operations. Large kernel size is preferred when salient information are globally distributed in the image. On the other hand, locally distributed information call for smaller kernels. These limitations pertaining to prevailing CNN architectures led to the development of a series of Inception networks. Szegedy et al. [41] proposed Inception v1 by stacking filters of various sizes on the same level, making the network wider rather than deeper. The outputs from different filters are concatenated and sent to the next layer. Additionally, 1×1 convolutions are introduced aiming at reducing the dimension of input channels and thereby reducing the computational cost. Szegedy et al. [35] proposed Inception v2 by implementing a set of iterative improvements over Inception v1. The authors factorized a 5×5 convolution layer to two 3×3 convolution layers, which reduced the computational cost significantly. The cost was further reduced by replacing a $n \times n$ convolution with a $1 \times n$ convolution followed by a $n \times 1$ convolution. Additionally, the filter banks were expanded to curtail the representational bottleneck (loss of information due to excessive reduction in dimension). More details about the Inception v2 network can be found in [35].

2.3.2 ResNet-50 and ResNet-101

Very deep neural networks are cursed with the problem of vanishing gradients. As a result, the performance of the network saturates and eventually starts degrading with increase in depth. He et al. [36] introduced the idea of 'skip connection' which enables the activation of one layer to be fed directly to another layer much deeper in the network bypassing one or many intermediate layers. This eventually led to the development of residual block which

facilitates training of very deep neural networks without any appreciable loss of performance. The objective in residual network is to ensure that a deep network does not produce higher training error than its shallower counterpart. Skip connection introduces an identity mapping which is easier for the residual block to learn pushing the residual function to zero. This ensures that addition of extra layers does not adversely impact the accuracy of the network, and the deep network performs at least at par with its shallower counterpart. On top of that, if the added layers manage to learn something useful then the deep residual network can even outperform its non-residual version. Exploiting this notion of residual block, a series of deep networks are developed. ResNet-50 and ResNet-101, having 50 and 101 convolutional layers respectively, are investigated in this study. The details of the ResNet architectures can be found in [36].

2.3.3 Inception-ResNet-v2

Inception-ResNet-v2 network [37] incorporates Residual connections proposed by He et al. [36] in combination with latest developments in the Inception architecture [35]. Residual connections add the convolution outputs of the Inception module to the input. Residual block requires that input to and output from the convolution module have the same dimension. 1×1 convolutions are used to compensate for the dimensionality reduction induced by the Inception blocks. The pooling operations inside the original inception modules were replaced in favor of the residual connections and the same was retained in the reduction blocks. The residual activations were scaled by a factor ranging from 0.1 to 0.3 to get rid of vanishing gradients. A detailed discussion on Inception-ResNet-v2 is beyond the scope of the present study and can be found elsewhere [37].

2.4 Datasets and Experimental Program

Images of buildings damaged by earthquakes experienced in the recent past in different parts of the world (Nepal, Taiwan, Ecuador, Erzincan, Duzce, Bingol, Peru, Wenchuan, and Haiti) are used in this study for evaluation of the proposed approach. The data were acquired from the Datacenterhub of Purdue University, USA [42]–[45]. Diversity in training

Figure 2.3. Illustrative examples of images depicting wide variation in lighting condition and data quality

data is necessary for reducing model variance, which is a measure of sensitivity of the model to specific observations. A learning algorithm with high variance performs well on training data. However, the performance declines when the model encounters data which is not used for training. The problem of high variance can be alleviated by introducing variations in the training data. Learning models trained with data collected from various sources are supposed to be more robust when tested on previously unseen data. The images used in this study for training of the detection algorithms represent wide variations in image resolution, lighting condition, blurring and degree of distortions. Specifically, the database comprises images with 69 different resolutions. The sample images presented in Figure 2.3 are illustrative of the wide-ranging lighting conditions encountered in the dataset. The database is enriched with diversity which resulted in better generalization capability of the learning models. All the damages observed under the said earthquakes were subdivided into four categories. The first damage category (Damage-1) denoted surface cracks. Spalling, which includes facade spalling and surface spalling, constituted the second damage category (Damage-2). The third damage category (Damage-3) was composed of spalling with exposed rebars. Severely buckled rebars formed the fourth damage category (Damage-4). Some example images representing the four damage categories are shown in Figure 4.2. The images were manually annotated and were divided into training set and validation set. 4-fold cross validation was conducted to examine how well the detection models generalize to independent datasets. The distribution of training and validation data at each round of cross-validation is shown in Table 2.1. Ten percent of all available data was used for validation at each cross-validation round, and the remaining 90% was used for training. No sample was used twice for validation. The evaluation metrics obtained from all 4 rounds of cross-validation were averaged to produce a single estimation. It is evident from Table 2.1 that the training data has uneven representation from different classes, which can potentially make the predictor biased towards the over-represented classes. To mitigate this problem of class imbalance, class specific weights are assigned to the loss function so as to impose additional penalty for misclassifying an under-represented class [38].

(a)

(b)

(c)

(d)

Figure 2.4. Damage categories considered for detection - (a) Damage-1: Surface crack, (b) Damage-2: Spalling, (c) Damage-3: Spalling with exposed rebars, (d) Damage-4: Severely buckled rebars.

Table 2.1. Category-wise sample size used for training and validation.

Cross-validation round	Damage 1		Damage 2		Damage 3		Damage 4	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation
1	865	97	1751	272	554	148	473	126
2	868	94	1791	232	547	155	437	162
3	773	189	1685	338	622	80	563	36
4	863	99	1811	212	479	223	490	109

2.5 Implementation Details

The Faster RCNN algorithm was implemented using TensorFlow open-source software library [46] and was run on two NVIDIA Titan X (Pascal) GPUs. Faster RCNN is designed to work with variable image size and aspect ratio. However, previous studies indicated that resizing the images enhances the performance. Therefore, the input images were resized in this study to a minimum dimension of 600 pixels and maximum dimension of 1024 pixels keeping the aspect ratio intact. In other words, if the longer dimension of the input image is less than 1024 pixels, then the shorter dimension is resized to 600 pixels, and the longer dimension is modified proportionally keeping the aspect ratio same. On the other hand, if the longer dimension of the input image is greater than 1024 pixels, then the longer dimension is resized to 1024 pixels, and the shorter dimension is resized appropriately to keep the aspect ratio unchanged. The input images were horizontally flipped randomly with a probability of 0.5 as part of the data augmentation. Then features are extracted from the input image using a sequence of convolutional layers which were a part of CNN architectures considered in this study. As for Inception-ResNet-v2, a set of atrous filters are slid over this feature map to carry out atrous convolution [47]. This enables object encoding at multiple scales by extending the receptive field without increasing the number of parameters and number of operations. In order to generate region proposals using RPN, Ren et al. [38] proposed 9 anchor boxes ($k = 9$) with 3 different scales and 3 different aspect ratios. However, a wider range of scales and aspect ratios leading to a higher value of k was found to enhance the detection accuracy significantly in this study. Various scales used in this study for anchor box generation include 0.125, 0.25, 0.5, 1.0, 2.0, 4.0 and 8.0. The aspect ratios had the values of 0.125, 0.5, 1.0, 2.0, 4.0, 6.0, 8.0 and 10.0. The minimum of input height and width

was considered as base anchor size. The anchor boxes were strided by 8 pixels both along the height and the width. Three hundred region proposals were generated per image calling for elimination of multiple detections. To filter all the duplicate boxes, a greedy procedure called non-maximum suppression (NMS) [48] was employed, where all candidate boxes are first sorted in the order of their objectness score. The best scoring box was selected and all other boxes having an intersection-over-union (IoU) greater than 0.7 with the selected box were discarded. IoU is an evaluation metric which is defined as the ratio of area of overlap to the area of the union between a ground-truth box and a predicted box. The remaining boxes were then classified and refined using a Fast RCNN module. The IoU threshold used for NMS at this stage was 0.6. The weights of the Inception-ResNet-v2 network were initialized by a model pretrained on MSCOCO dataset and fine-tuned thereon. MSCOCO [49] is a large repository of images (328k) containing 90 different objects that are commonly encountered in everyday life. The said model had 90 neurons in the last layer representing 90 classes. Therefore, this layer was replaced by one with only 4 neurons in this study. The weights for last layer was initialized from a uniform distribution as suggested by Glorot and Bengio [50]. All the weights are subsequently updated using Stochastic Gradient Descent (SGD) [51] with a momentum value of 0.9. The problem of exploding gradient is commonly encountered while training a very large neural network. The gradients shoot off exponentially during successive back propagation through the network layers, rendering the learning process highly unstable. This problem can be averted by clipping the gradients by a preset threshold. A threshold of 10 was set for the gradient norm to this end. The initial learning rate was set to 0.003 and was gradually reduced thereafter with training steps.

2.6 Results and Discussions

The performance of the proposed approach was evaluated on the validation data described in Section 2.4. The test images were input to the trained network and bounding boxes were predicted with respective object class as shown in Figure 2.6. The predicted boxes were compared with the ground truth boxes, and any prediction having an IoU greater than a threshold of 0.5 was considered to be *true positive*. If multiple boxes are predicted for a

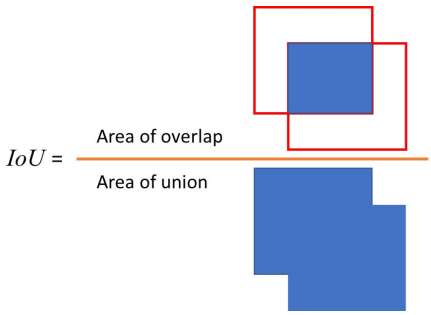
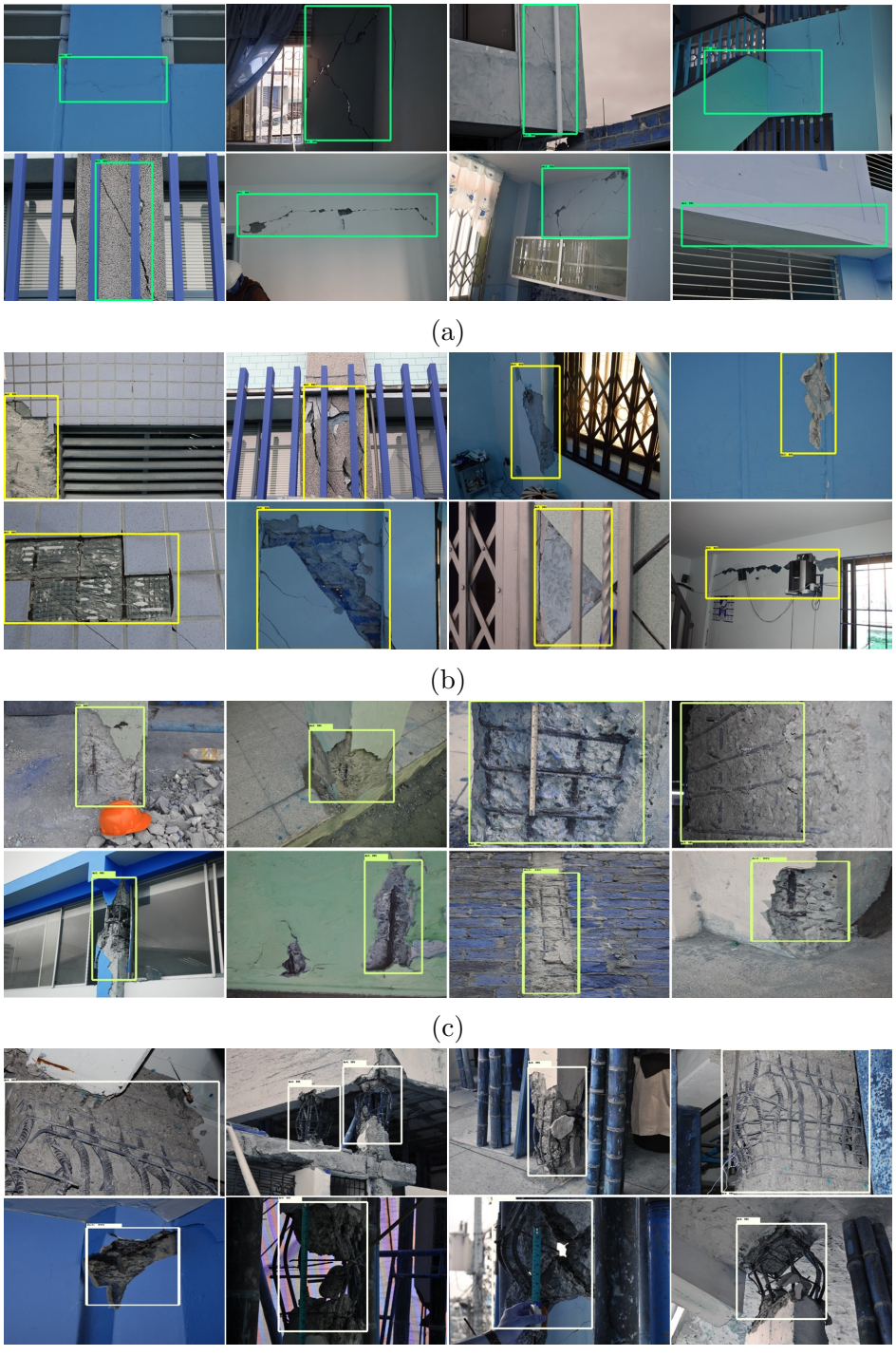


Figure 2.5. Intersection over Union (IoU): It is the ratio of area of intersection to the area of overlap between two boxes.



(a)

(b)

(c)

(d)

Figure 2.6. Sample detection results - (a) Damage-1: Surface crack, (b) Damage-2: Spalling, (c) Damage-3: Spalling with exposed rebars, (d) Damage-4: Severely buckled rebars.

single ground truth box, then only the highest scoring box is considered to be *true positive* and rest all are dubbed as *false positive*. If a ground truth box does not possess any predicted box associated with it, it is designated as *false negative*. The detection performance of the proposed algorithm with four different CNN architectures are measured in terms of precision and recall values. Precision is defined as the ratio of *true positive* to the sum total of *true positive* and *false positive*. In other words, it tells us what percentage of the overall detections are correct detections. The mean and standard deviation of precision values obtained from four rounds of cross-validation are reported in Table 2.2. For instance, the precision values obtained from Inception-ResNet-v2 architecture for four classes considered in this study had the mean values of 65.5%, 50.0%, 52.0%, and 53.8%, respectively, while the corresponding standard deviation values were estimated as 7.9%, 9.5%, 5.6% and 5.5%, respectively. It means that 65.5% of all predicted boxes classified as Damage-1 belong to the correct detections in average sense. Similar interpretations can likewise be extended to other damage classes and CNN architectures. Another evaluation metric which is often used alongside precision score is recall. It is the ratio of *true positive* to the sum total of *true positive* and *false negative*. It indicates what percentage of the actual ground truth objects have been successfully identified by the detection network. The mean recall values that the trained network produced for four classes with Inception-ResNet-v2 architecture were 78.8%, 65.3%, 62.5%, and 59.5%, respectively (Table 2.3). Corresponding standard deviations were evaluated as 8.2%, 16.0%, 4.2%, and 7.8%, respectively. In other words, 78.8% of all damages annotated as Damage-1 were correctly predicted on the average, and likewise for other classes and CNN architectures. The said precision and recall values are considerably higher in comparison to that reported by Yeum et al. [30] for single class (spalling) detection (Precision: 40.48%, Recall: 62.16 %) on similar dataset. Minor cracks in concrete are typically hard to detect due to potential noise infusion [32]. However, the earthquake induced cracks used in this study were, by and large, distinct and easily detectable. On the other hand, the other three damage categories were all related to spalling and therefore contained significant visual correlation, which made them less distinguishable from each other resulting in a lot of classification error. This potentially led to relatively high detection performance vis-à-vis damage-1 as compared to rest of the damage categories.

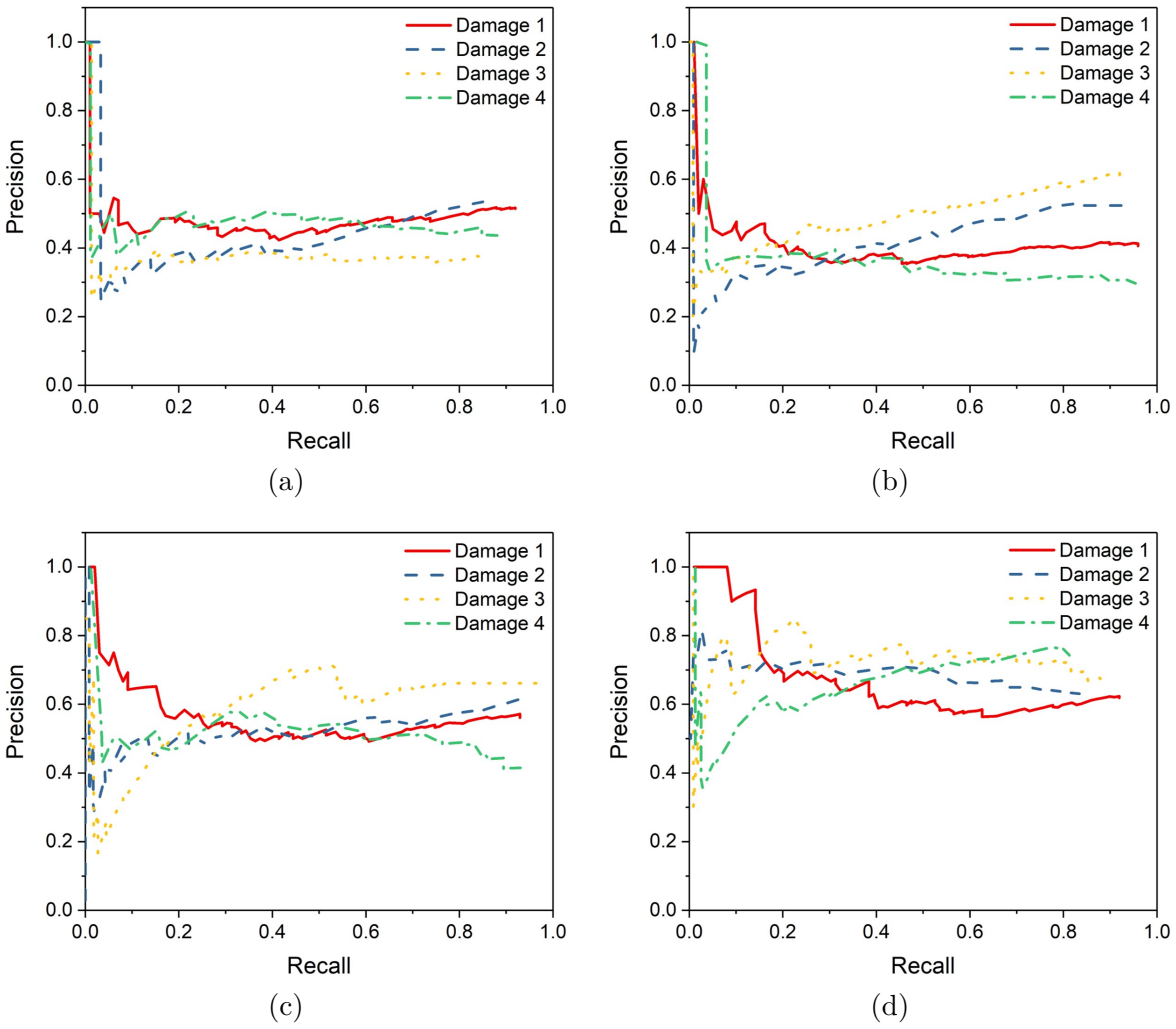


Figure 2.7. Precision - recall curves - (a) Inception v2, (b) ResNet-50, (c) ResNet-101, and (d) Inception ResNet v2.

Precision and recall values are inversely related. Setting the detection threshold to a low value will allow the network to predict most of the objects in the image. However, it will generate a large number of false positives at the same time. On the other hand, a high value of the detection threshold will produce very few false positive. However, it will result in numerous missed detections. It is therefore customary not to rely entirely on either of the two decision metrics for the sake of comparison among different detection models. Alternatively, the entire precision-recall curve (Figure 2.7) is looked into and the area under the curve is used as an evaluation metric. This parameter, also known as the average precision (AP), sums up the precision-recall curve to a single number. Higher values of AP indicate better performance of the detector. The AP values are calculated from the precision-recall curves for all four damage types and all four CNN architectures considered in this study and the mean and standard deviation values (averaged over four rounds of cross-validation) are reported in Table 2.4. Figure 2.8 presents more detailed information with regard to the dispersal of all the evaluation metrics (precision, recall and AP) for different CNN architectures and damage categories. The variation range formed by one standard deviation on either side of the mean value is represented by a rectangular box in this figure.

A careful analysis of the information presented in Table 2.4 indicates that no consistent pattern exists in the performance hierarchy of the four CNN architectures evaluated with respect to the mean AP value. For instance, ResNet-50 produced a higher mean AP for Damage-1, Damage-2 and Damage-3 in comparison with the Inception v2 architecture. However, the trend was reversed for Damage-4 where Inception v2 outperformed the ResNet-50 architecture in terms of the same evaluation metric. Similarly, ResNet-101 performs better than ResNet-50 in terms of mean AP in identifying Damage-1 and Damage-4. However, when it comes to the detection of Damage-2 and Damage-3, ResNet-50 turns out to be more efficient. This anomaly can be resolved by averaging the AP values over all object classes. The evaluation metric thus generated is called mean average precision (MAP), which is typically used to compare the efficiency of different detection algorithms. The mean MAP values (averaged over four rounds of cross-validation) for all four CNN architectures are shown in Figure 2.9. It is evident from the figure that Inception v2 architecture afforded the lowest accuracy with a mean MAP value of 51.0%. A 3.0% increase in the mean MAP value was

Table 2.2. Mean (μ) and standard deviation (σ) of precision for different CNN architectures

Architecture	Damage 1		Damage 2		Damage 3		Damage 4	
	μ	σ	μ	σ	μ	σ	μ	σ
Inception v2	0.576	0.110	0.432	0.079	0.410	0.061	0.497	0.092
ResNet-50	0.532	0.121	0.405	0.091	0.448	0.068	0.403	0.111
ResNet-101	0.613	0.089	0.435	0.096	0.423	0.045	0.458	0.103
Inception ResNet v2	0.655	0.079	0.500	0.095	0.520	0.056	0.538	0.055

Table 2.3. Mean (μ) and standard deviation (σ) of recall for different CNN architectures

Architecture	Damage 1		Damage 2		Damage 3		Damage 4	
	μ	σ	μ	σ	μ	σ	μ	σ
Inception v2	0.750	0.112	0.628	0.118	0.553	0.043	0.507	0.116
ResNet-50	0.750	0.153	0.598	0.125	0.572	0.068	0.545	0.087
ResNet-101	0.770	0.108	0.638	0.131	0.565	0.040	0.578	0.098
Inception ResNet v2	0.788	0.082	0.653	0.160	0.625	0.042	0.595	0.078

Table 2.4. Mean (μ) and standard deviation (σ) of AP for different CNN architectures

Architecture	Damage 1		Damage 2		Damage 3		Damage 4	
	μ	σ	μ	σ	μ	σ	μ	σ
Inception v2	0.566	0.096	0.455	0.028	0.490	0.090	0.528	0.023
ResNet-50	0.577	0.102	0.505	0.036	0.568	0.057	0.508	0.113
ResNet-101	0.658	0.107	0.494	0.056	0.529	0.094	0.538	0.131
Inception ResNet v2	0.681	0.080	0.554	0.068	0.627	0.089	0.570	0.088

observed when ResNet-50 was employed. Invoking ResNet-101 rendered a further increase of 1.5% to the same. However, the best performance was observed with Inception-ResNet-v2 architecture which produced a mean MAP value of 60.8%.

Apart from accuracy, another parameter which is often taken into account while comparing various detection algorithms is the computational cost. The computational cost is measured in this study in terms of average processing time for a single image. It was observed that the architectures that exhibited higher MAP values actually had slower processing speed (Figure 2.10). This led to the conclusion that detection accuracy and processing speed are inversely related, and the selection of a suitable detector is a trade-off between the two.

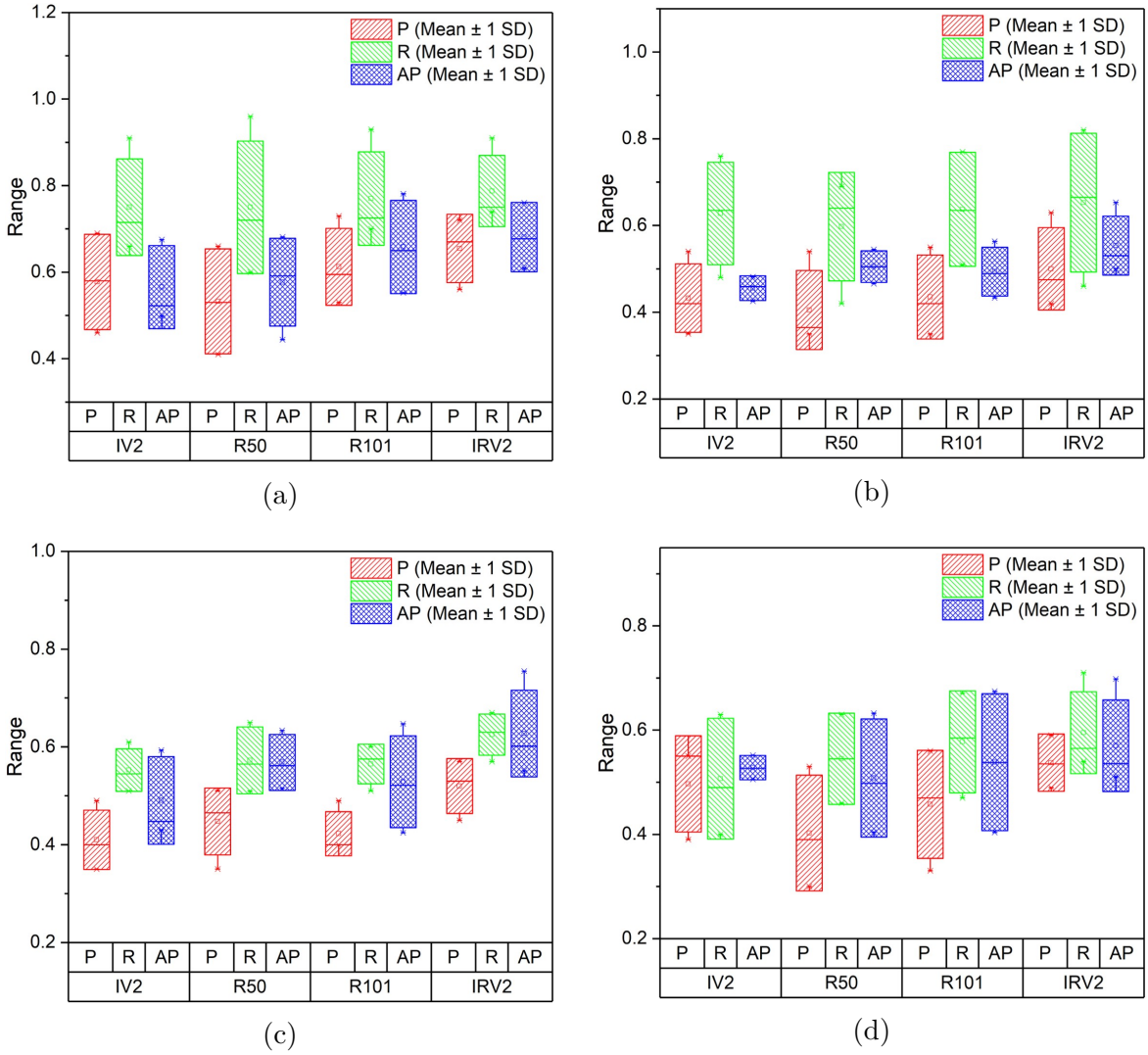


Figure 2.8. Variation of evaluation metrics over all rounds of cross-validation for - (a) Damage-1, (b) Damage-2, (c) Damage-3, and (d) Damage-4. IV2: Inception v2, R50: ResNet-50, R101: ResNet-101, IRV2: Inception ResNet v2, P: Precision, R: Recall, AP: Average Precision, SD: Standard deviation.

However, the values presented in Figure 2.10 are highly subjective and are dependent on the image resolution and specific GPU architecture used, and therefore, should not be taken in the absolute sense. However, in relative terms, it can be inferred that Inception v2 is the fastest of all architectures considered in this study. ResNet-50 and ResNet-101 take about 1.2 and 1.7 times the time taken by an Inception v2 architecture to process an image of the same resolution. However, Inception ResNet v2 was identified as the slowest of all considered architectures taking about 5.6 times the time taken by Inception v2 to accomplish the same task.

The ultimate objective of developing damage detection algorithm is to integrate it with robotic systems for autonomous inspection. A major challenge that is encountered to this end is wide-ranging camera specifications leading to huge variations in image resolution and quality which may potentially affect the performance of the proposed neural network-based approach. However, it should be noted here that the images used in this study for training and validation of the neural networks were collected from nine different past earthquakes, and the resulting datasets contained huge variations in image resolution, lighting condition, blurring and other distortions. This enriched the database with diversity and made the neural network robust against previously unseen data and also added to its generalization ability. Relatively large dispersal in the evaluation metrics as observed in Figure 2.8 and Tables 2.2-2.4 are a direct consequence of this diversity. Damage detection on a number of images captured by an UAV-mounted camera in the aftermath of Taiwan earthquake (2016) is presented in Figure 2.11. The damages were detected by the trained Faster RCNN algorithm with Inception-ResNet-v2 as backbone architecture. Limited computation capability of on-board processing units is another bottleneck in robot-based real-time damage diagnosis. Commercially available portable power-efficient embedded AI computing devices such as NVIDIA Jetson TX2 can provide a viable solution to this problem and is a scope for future research. Future studies should also focus on making the network smaller, faster and consequently more suitable for on-board real-time computation by pruning of redundant neurons which do not contribute significantly to the network outputs, as demonstrated by [52].

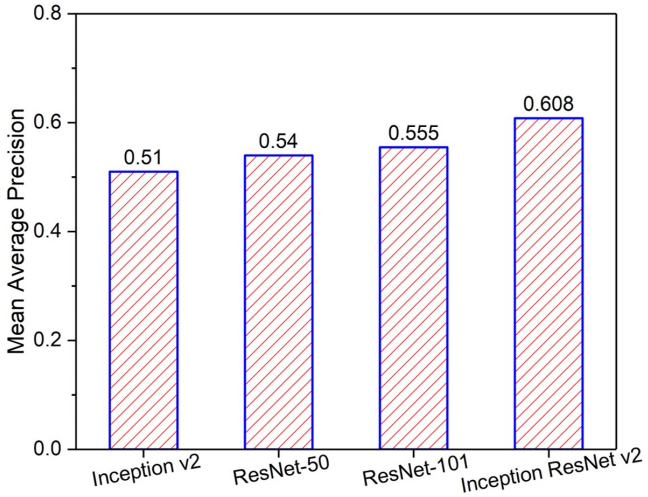


Figure 2.9. Comparison of mean average precision (MAP) for different CNN architectures.

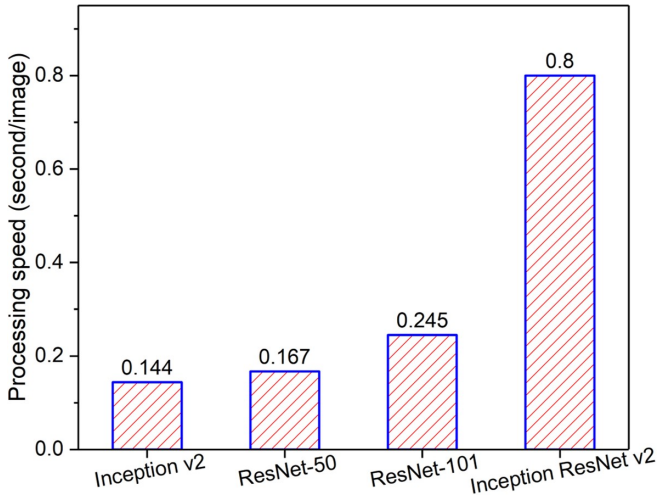


Figure 2.10. Comparison of processing speed for different CNN architectures.

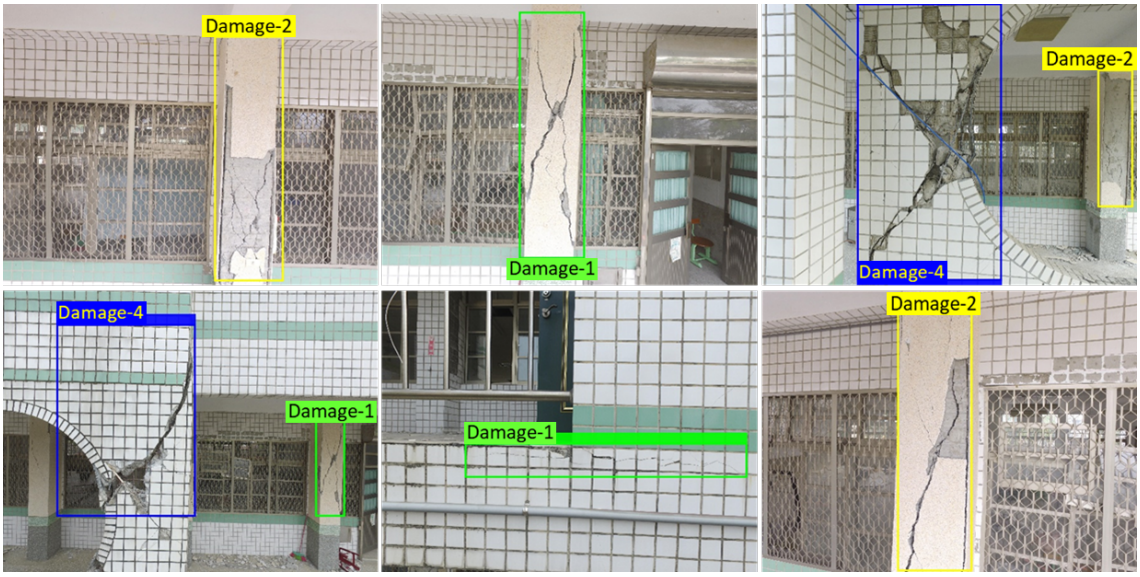


Figure 2.11. Damage detection results for images captured by a UAV post Taiwan earthquake (2016). Predicted boxes for Damage-1 (surface crack), Damage-2 (spalling), Damage-3 (spalling with exposed rebars), and Damage-4 (severely buckled rebars) are shown in different colors.

2.7 Conclusions

Faster RCNN algorithm is used in this study to detect multiple damage categories in reinforced concrete buildings. Four different CNN architectures, namely, Inception v2, ResNet-50, ResNet-101 and Inception-ResNet-v2 are exploited to this end. A pretrained model was used for initialization of the network weights which were subsequently fine-tuned by stochastic gradient descent optimization approach with momentum. The networks were trained using image data collected from several past earthquakes, namely, Nepal (2015), Taiwan (2016), Ecuador (2016), Erzincan (1992), Duzce (1999), Bingol (2003), Peru (2007), Wenchuan (2008), and Haiti (2010). Four different damage categories were considered in this study, namely, surface crack, spalling, spalling with exposed rebars, and severely buckled rebars. The performance of the trained networks was evaluated on the validation dataset. It was observed that Inception-ResNet-v2 significantly outperforms the other networks considered in this study producing a MAP value of 60.8%. It was also noted that processing speed of the detection algorithms reduces with increase in accuracy. It is believed that this study will broaden the scope for vision based autonomous inspection of civil infrastructures.

3. AUTONOMOUS VISION-BASED DAMAGE CHRONOLOGY FOR SPATIOTEMPORAL CONDITION ASSESSMENT OF CIVIL INFRASTRUCTURE USING UNMANNED AERIAL VEHICLE

3.1 Background

This chapter[1]presents a novel comprehensive approach to health monitoring of civil infrastructures by introducing a time dimension into the vision-based condition assessment pipeline. It is shown that useful information can be extracted from an archive of inspection images by employing computer vision-based algorithms. Identification of a damage during the course of a recent inspection initiates an exhaustive search into the historical data collected during the previous rounds of inspection. Corresponding images are identified and synthesized to generate a reconstructed view of the scene pertaining to each inspection round. Regions of interest are subsequently extracted from the reconstructed scenes leveraging a CNN-based detection model. This is followed by damage segmentation and quantification exploiting state-of-the-art morphological and image processing techniques paving the way for time-based evaluation of damage severity and cognizant decision making. The methodology presented in this work is robust against noise intrusion and changes in illumination condition. It does not assume any prior knowledge about damage locations and provides a great deal of flexibility with regard to camera poses and orientations. The proposed approach can be applied to data collected by human inspectors using hand-held cameras (e.g. smartphone camera). However, it is most appropriate for autonomous inspection assisted by vision systems mounted on mobile robots including UAVs. Cracks on concrete surface is used as a case study to demonstrate the feasibility of this approach. However, it can be extended to other defect categories such as spalling and corrosion, with appropriate modifications.

[1]The content of this chapter is published as follows: T. G. Mondal and M. R. Jahanshahi, "Autonomous vision–based damage chronology for spatiotemporal condition assessment of civil infrastructure using unmanned aerial vehicle," *Smart Structures and Systems*, 25(6), 733-749, 2020, https://doi.org/10.12989/sss.2020.25.6.733

3.1.1 Related Works

A number of studies in the past explored time-based evaluation of structural defects. Digital image correlation (DIC) is exploited by many researchers ghorbani2015full in the past to measure full-field displacement and strain. However, this technique relies on static camera, and therefore can only be used in situations where the damage location is known a priori. Moreover, it necessitates painting of speckle patterns on the structure under investigation to produce distinct visual features, which is not feasible in large structures like buildings and bridges and in situations where the surface of the structure is physically inaccessible. The approach presented in the current study is free from all such limitations. A movable (hand-held) SLR camera and a camera mounted on an unmanned aerial vehicle (UAV) are used for data collection which eliminated the need of prior knowledge about damage locations. Besides, this method is contactless in true sense of the term, as it does not require any speckle pattern to be pained on the surface to be inspected. Kong and Li [53] used image overlapping technique to detect fatigue cracks in civil infrastructures. The authors relied on differential image features engendered by crack breathing as indicators for crack identification. This approach will inherently fail when the background undergoes visual changes over time owing to accumulation of dirt, rust, stains, etc. Moreover, it requires all images to be captured from similar camera poses and lighting conditions, which imposes a serious constraint on extensive use of this technique. Detection of fatigue crack in steel bridges was also studied by Kong and Li [54] exploiting video-based feature tracking. Movement of each feature was tracked through a video stream and the presence of a crack was indicated by differential movement pattern exhibited by the feature points inside a localized circular region. However, optical flow-based feature tracking process is fraught with many limitations. It assumes brightness constancy. So, it does not perform well in situations that involve change in illumination condition. It cannot effectively deal with scale variation and viewpoint changes. Besides, it works well only under small displacements, and therefore not suitable for large motion. These limitations are dispelled in the current study in many ways. The approach presented in this study is scale invariant, and robust against noise intrusion and changes in illumination condition. It can detect large motion. Moreover, it affords the flexibility of capturing

images from varied camera positions and orientations, which is a major advantage of this approach. Jahanshahi et al. [55] proposed a vision-based approach for estimating damage evolution through multi-image stitching and scene reconstruction. However, the camera was constrained in this study in regard to translation. In the present study, this constraint is relaxed enabling the camera to rotate and translate without any restrain. Besides, the approach presented in this study [55] is not fully autonomous in the sense that a human inspector needs to compare the current scene with its previous condition and deduce the damage evolution manually. In other words, the proposed technique relied on inspector's judgment vis-à-vis evolution of the damage. This makes the entire procedure tedious, labour-intensive, subjective and qualitative. The present study addresses this limitation by including an autonomous localization and quantification module in the damage diagnosis pipeline making the entire process faster and more efficient. Additionally, quantitative and time-based evaluation of damage severity makes it possible to predict residual life of a structure and to take precautionary measures, if necessary.

3.1.2 Scope

The remaining of the chapter is arranged in the following order. Section 3.2.1 presents an overview of the test protocol and data collection procedure. Section 3.2.2 deals with various components of the correspondence identification technique adopted in this study. The details of damage detection approach are presented in Section 3.2.3. The necessary theoretical background for damage quantification is presented in Section 3.2.4. The results are presented and discussed in Section 3.3. Finally, conclusions are summarized in Section 3.4.

3.2 Methodology

As a case study, cracks on concrete surface are used to illustrate the nuts and bolts of this approach. A reinforced concrete beam was tested in the laboratory subjecting it to a gradual load increment in order to simulate a progressive damage. Cracks appearing on the beam surface were photographed after every stage of load increment and the images were

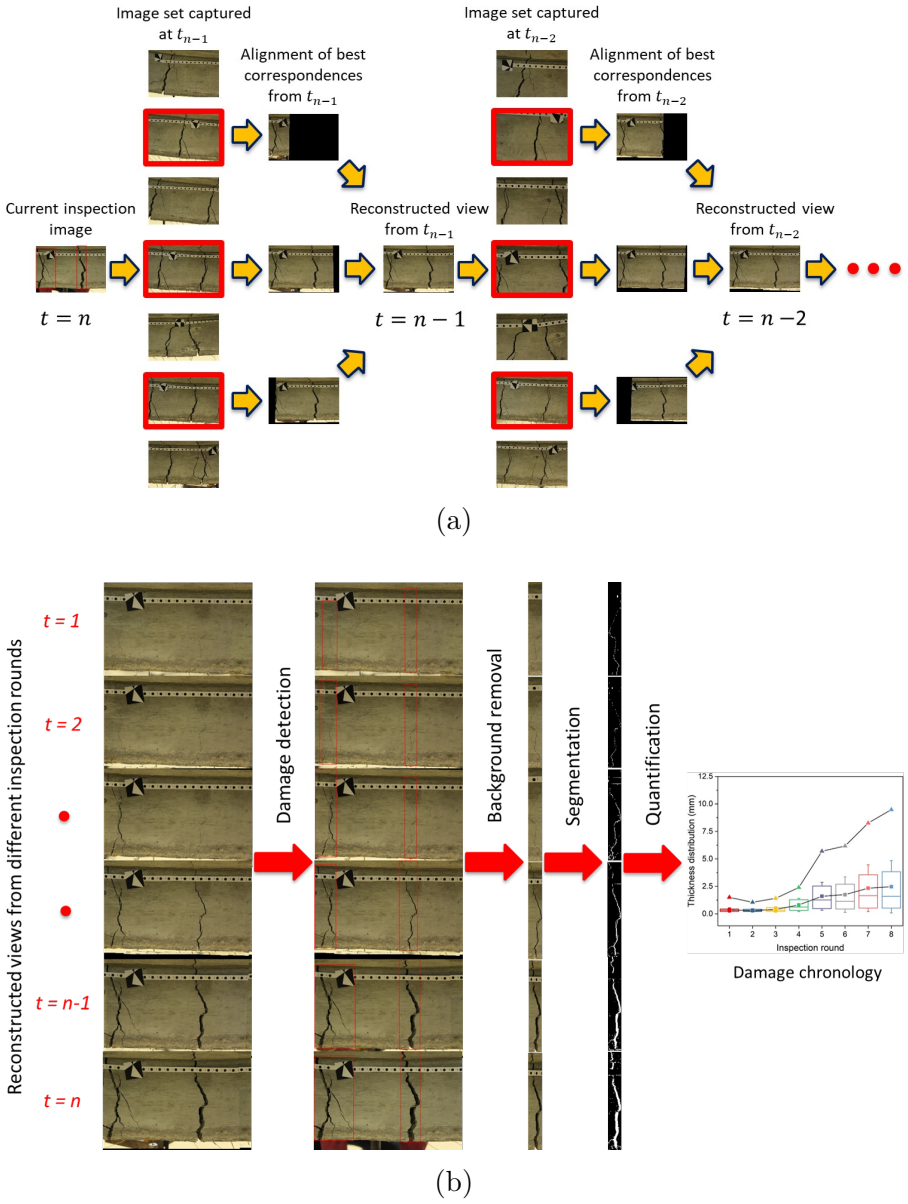
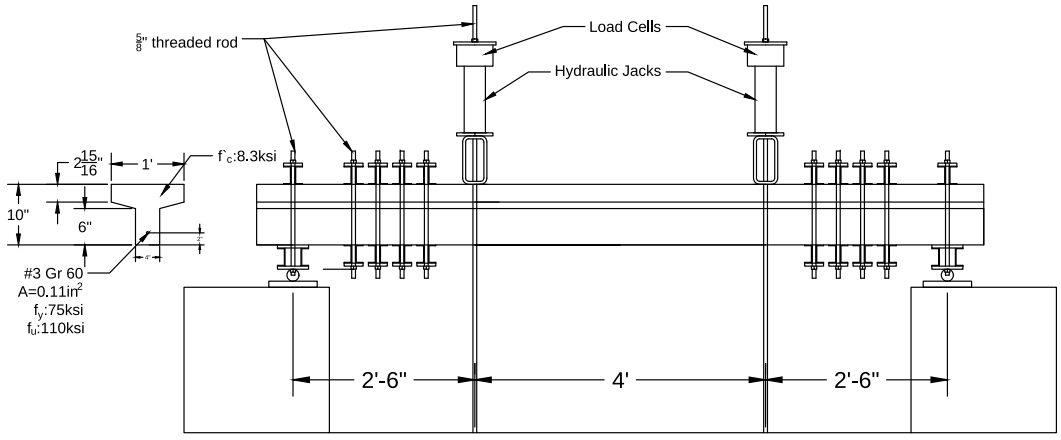
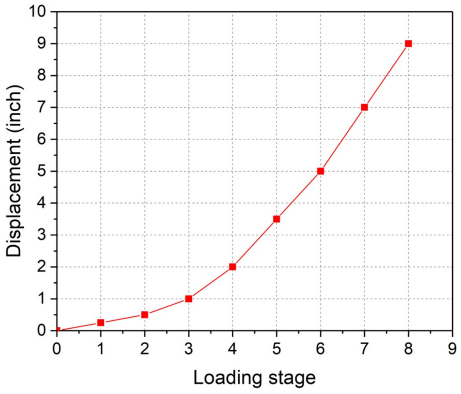


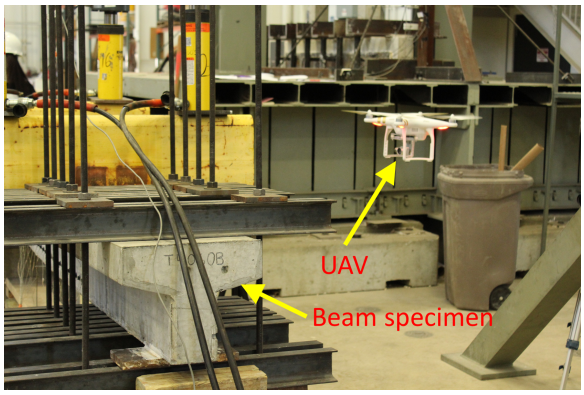
Figure 3.1. The layout of the proposed approach - (a) Correspondence identification from the preceding data set based on spatial proximity, registration of the best correspondences onto the plane of the current reference image, and repetition of the same procedure over all previous data sets to generate a temporally ordered set of 2D reconstructions of the concerned damaged area. (b) Detection of damage on the reconstructed views from the past, extraction of interest area to remove nonessential background, damage segmentation, followed by quantification and time-based visualization.



(a)

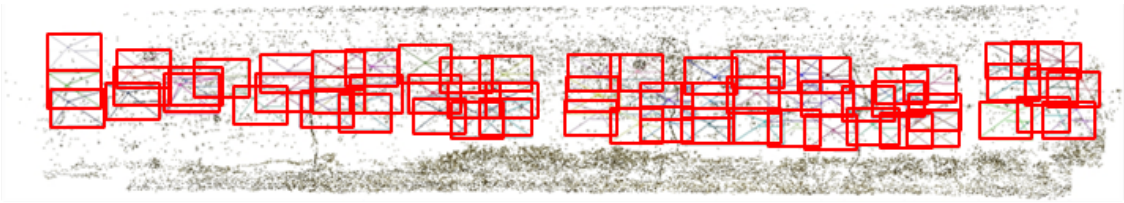


(b)

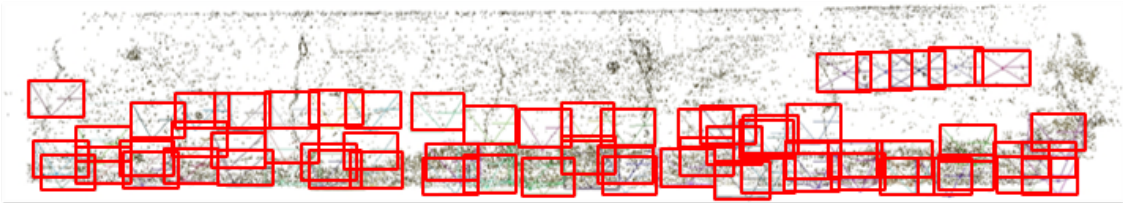


(c)

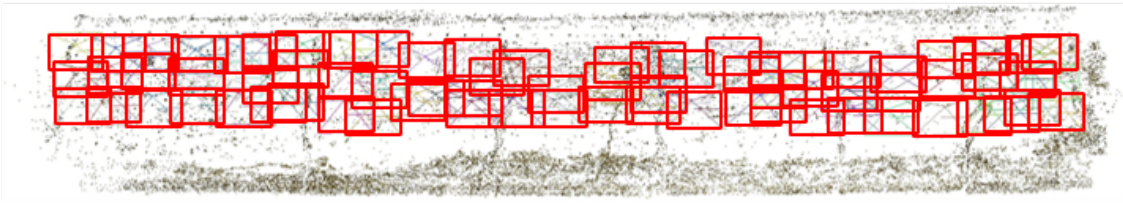
Figure 3.2. Experimental setup for data collection - (a) Experimental setup, (b) Loading protocol, (c) Data collection.



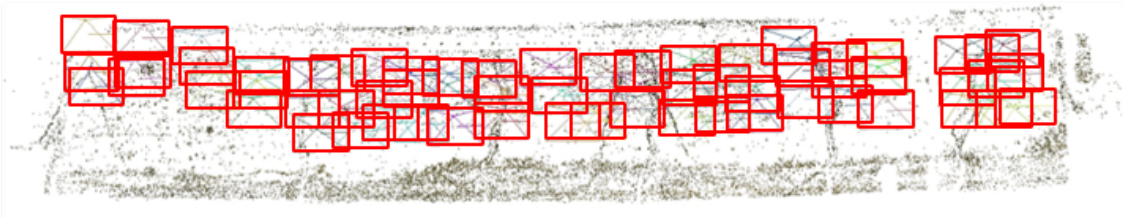
(a)



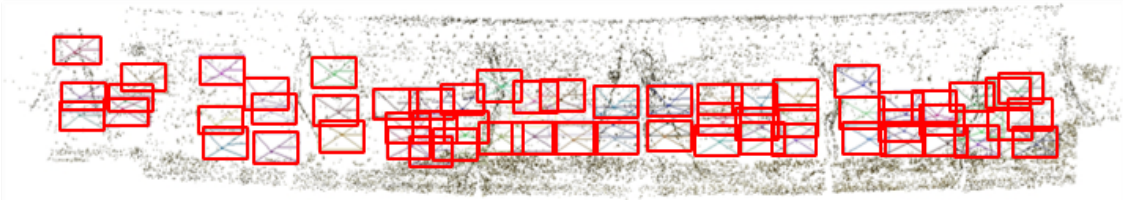
(b)



(c)



(d)



(e)

Figure 3.3. Data collection path of a hand-held SLR camera for different inspection rounds - (a) round-1, (b) round-2, (c) round-3, (d) round-4, and (e) round-5. The prisms denote the camera poses and orientations, and the point clouds denote the 3D scene reconstructions of the beam for each inspection round. It should be noted that the data collection path was not constant and it varied over inspection round.

time-stamped and saved in specific folders. This was accomplished by a hand-held SLR camera and a camera mounted on a UAV to emulate actual robot-based data collection where camera positions and orientations are not fully controllable. After the final round of load increment, the entire data set representing time evolution of concrete cracks were available for further analysis. Figure 3.1 presents an overview of the proposed algorithm. Identification of damage in the most current data set engenders an exhaustive search in the immediately preceding image set looking for correspondences. Speeded up robust features (SURF) [56] algorithm is used to identify interest points in the current inspection image and also in every single image in the previous data set. Feature matching is carried out based on Euclidean distance between two descriptor vectors and the candidates with large number of matched features are designated as potential correspondences. Homography transformation is computed for each selected correspondence through linear least square method and subsequent nonlinear refinement using Levenberg-Marquardt algorithm [57], [58], which is followed by registration of the corresponding images onto the plane of the current reference image. The warped images are then stitched to form a complete 2D reconstructed view of the concerned damage region from the immediately preceding data set. This procedure is repeated for all the previous data sets captured at different points in time considering the reconstructed view from the immediately succeeding data set as the reference. Temporally ordered set of 2D reconstructions thus produced chronicles the evolution of a damage in a manner conducive to time-based reasoning and lucid visual interpretation, and forms the basis for the next stage of the proposed algorithm, namely, damage identification and quantification. A notable detection algorithm called Faster RCNN [38] is leveraged to this end to localize the cracked area in the reconstructed images. The relevant portion of the images containing the cracks are then cropped out to get rid of the remaining nonessential background (undamaged), inclusion of which may have debilitating effect on the performance of the subsequent segmentation and quantification processes due to noise infusion. The cropped pixels are then segmented using a morphological approach, forming the basis for crack thickness quantification using distance transform method ([59], [60]). The approach presented in this study can be extended to other defect categories such as spalling and corrosion, with appropriate modifications. Availability of an inspection database which is complete in terms of coverage

of the damaged areas and that affords adequate overlap with adjacent images is a prerequisite for this approach. Besides, the algorithm may not perform well in absence of adequate visual features in the inspection images. In such situations, IMU and GPS information can be exploited for accurate scene reconstruction which is a scope for future research.

3.2.1 Experimental Setup and Data Collection

The database required for validation of the proposed approach was generated by testing a reinforced concrete T-beam in the laboratory under gradually increasing load in four-point bending configuration as shown in Figure 3.2a. The beam was tested in displacement control mode, and the applied displacement is shown in Figure 3.2b as a function of loading step. After every step of displacement increment, an intermission was appropriated during which the entire span of the beam was photographed using a hand-held SLR camera as well as a camera mounted on a UAV to capture the cracks that appeared on the surface (Figure 3.2c). The SLR camera was displaced laterally to photograph different segments along the span and depth of the beam ensuring adequate overlap between successive images (Figure 3.3). The camera movement was not controlled, and the data collection path varied over inspection round as evident from Figure 3.3. An archive of time-stamped images representing various levels of degradation was thus produced mimicking time-evolution of damage in concrete structures. This data set formed the basis for subsequent analyses which are described in the following sections.

3.2.2 Correspondence Detection and Alignment

Theoretical formulation of this algorithm presupposes the availability of a comprehensive visual data set built perennially through collection of images over several rounds of routine inspection by a human inspector or by an inspection robot ([61]–[63]). If a defect is detected during the inspection of a structure, it becomes necessary to know the history of evolution of the defect. That necessitates probing in to the data collected during previous rounds of inspection. The first challenge that is confronted to this end is identifying the relevant images corresponding to the defective region from a large database of archival images [64],

[65]. This can be achieved through a sequence of widely used computer vision algorithms such as feature detection, feature matching and image registration, as explained in the following sections.

Feature Detection

The first step in the correspondence identification pipeline is the detection of features or interest points (Figure 3.4b). Features are unique patterns which can be easily tracked and compared across several images. There are a number of techniques available in literature for detecting interest points in images. SURF algorithm is one such technique which is leveraged in this study. This algorithm locates high-variance interest points in an image which are invariant to scale, viewpoint and illumination changes. A local dominant direction is associated with each interest point and a 64 element normalized descriptor vector representing the local gray level variations with respect to the dominant direction is computed at each such point. The reader may refer to the original paper by Bay et al. [56] for more detailed discussion about this algorithm.

Feature Matching

Feature detection is followed by feature matching (Figure 3.4c), the objective of which is to identify the best match for a feature in one image from all the features in another image. Number of matched features is an indication of degree of resemblance between two images. Brute-Force matcher is used in this study, where the Euclidean distance between two descriptor vectors is used for similarity comparison. Two best matches are drawn for each feature in the first image. On occasion, the second best match is found to be very close to the best match owing to noise or other reasons. Such anomalies are tackled by computing the ratio of the closest distance to the second closest distance, and discarding all matches where this ratio is greater than 0.75 as suggested by Lowe [66]. This eliminates 95% of the false matches as shown in Figure 3.5. However, a small number of outliers are retained at this stage.

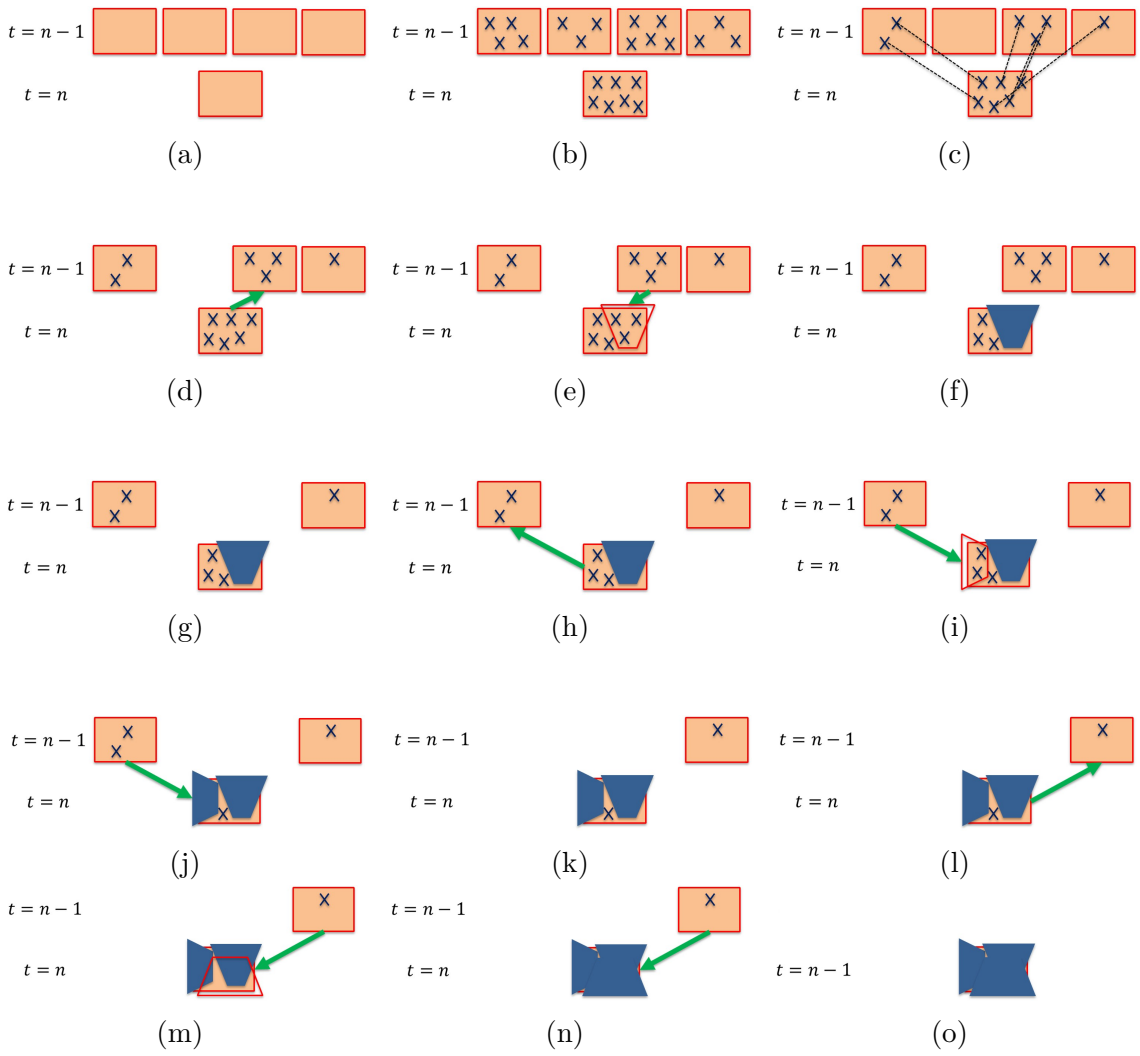
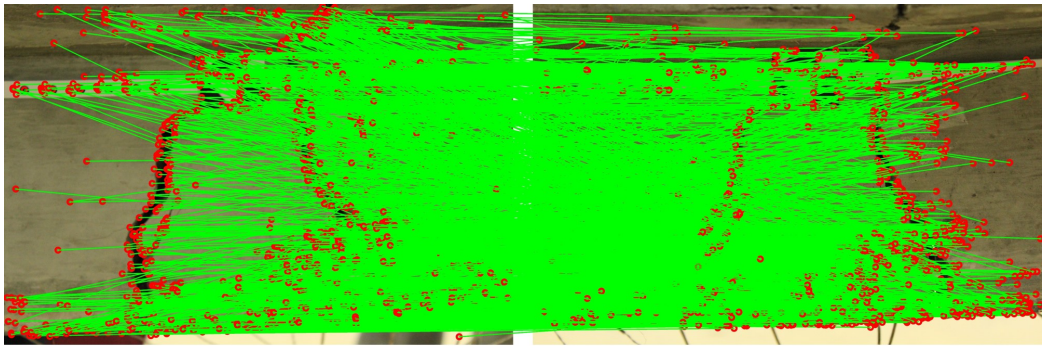
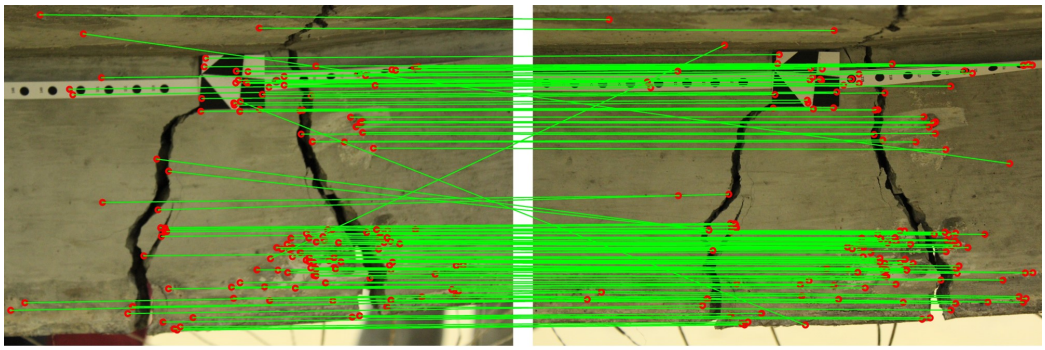


Figure 3.4. Illustrative diagrams outlining the steps for generating reconstructed view from previous data set: (a) Raw data, (b) Feature detection, (c) Feature matching, (d) Best correspondence, (e) Warping, (f) Registration, (g) Revised search region, (h) Next best correspondence, (i) Warping, (j) Registration, (k) Revised search region, (l) Next best correspondence, (m) Warping, (n) Registration, and, (o) Final reconstruction.



t_n t_{n-1}

(a)



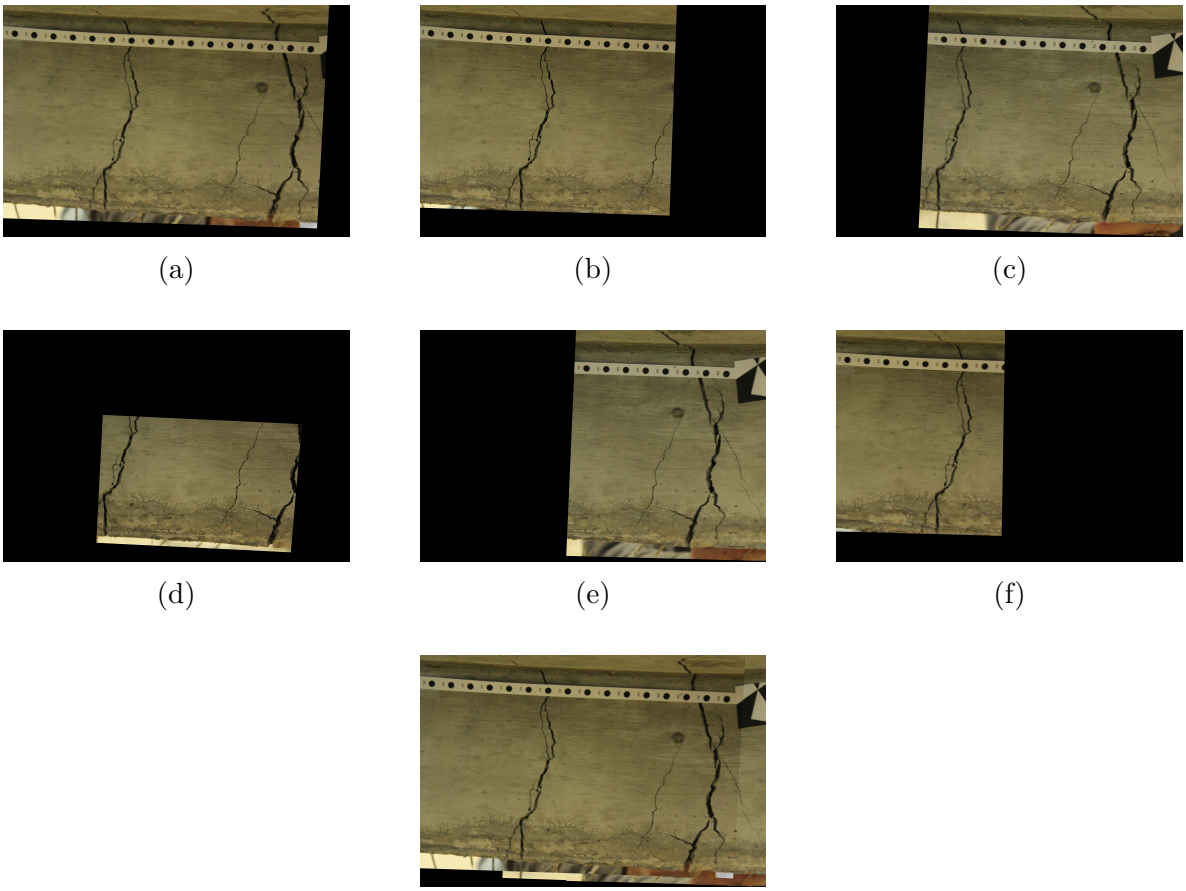
t_n t_{n-1}

(b)

Figure 3.5. Feature detection and matching: (a) Initially matched features, (b) Matched features after applying Lowe's ratio test.

Image Registration

The image in the immediately previous data set having the largest number of matched features vis-à-vis the reference image in the current data set is designated as the best correspondence (Figure 3.4d). Damage chronology can only be established when the corresponding images from previous data sets are aligned to the plane of the reference image. This requires estimation of the homography matrix between the reference image and the correspondences in the immediately preceding data set. It may be noted here that the homography is a 3×3 transformation matrix which maps the points in one image to the corresponding points in another image. Linear least square method is exploited in combination with an outlier rejection algorithm called RANdom SAmple Consensus (RANSAC) [67] to obtain an initial estimate for the homography matrix. This is followed by a nonlinear refinement of the estimated homography matrix using Levenberg-Marquardt algorithm based on the inlier points alone. The estimated homography matrix is then used to warp (Figure 3.4e) and register (Figure 3.4f) the best correspondence on the plane of the reference image. Following this, the matched features corresponding to the best correspondence are eliminated from the list of available features for the reference image (Figure 3.4g), and the next best correspondence is determined based on the revised list of matched features (Figure 3.4h). This next best correspondence is then registered on the plane of the reference image (Figures. 3.4i and 3.4j) in a similar fashion following the same procedure mentioned previously in this section. This process of correspondence identification and alignment is continued until the number of residual matched points corresponding to the reference image drops below a predefined threshold (100 in this study) or the number of identified correspondences reaches a preset value (which is set to 10 in this study). Upon completion of this process, all the warped correspondences are stitched together producing a complete 2D reconstruction (Figure 3.4l) depicting the prior condition of the scene in the reference image (Figure 3.6). The reconstructed view acts as a reference image for the next round of iteration, where the correspondences are identified from the immediately preceding data set. Eventually, an ordered set of reconstructed views are obtained portraying the evolution of a scene through time. One round of correspondence identification and alignment takes six minutes of processing time approximately.



(a) (b) (c)

(d) (e) (f)

(g)

Figure 3.6. (a)-(f) Warping and registration of correspondences, (g) View synthesis producing complete 2D reconstruction.

3.2.3 Damage Detection

This section describes the process for autonomous segmentation of damages in the chronologically ordered reconstructed views of the scene under consideration. Previous studies primarily focused on two approaches for detecting cracks in images, namely, edge-based techniques and morphological techniques. Jahanshahi et al. [68] compared the pros and cons of the two approaches and concluded that morphological techniques outperform edge-based techniques in presence of non-crack edges. Therefore, morphological approach is adopted in this study for extracting cracks from the images. However, presence of surface irregularities may produce false positives leading to inaccurate segmentation [8]. This can be averted by secluding the damaged region of interest from the remaining image. Deep learning-based approaches have been used by several researchers in the past [6], [7], [69] to localize defects in images. This study leverages Faster RCNN algorithm to this end. This eliminates a large part of the nonessential background significantly diminishing the scope of noise infusion in the morphology-based segmentation process.

Damage Localization using Faster RCNN

In Faster RCNN, a CNN is first used to generate a feature map from the input image. Inception-ResNet-v2 network [37], which incorporates Residual connections [36] and Inception module [35], is used to this end in this study. Thereafter, Region Proposal Network (RPN) [38] is used to generate region proposals. RPN is a fully convolutional network trained to predict object bounds and objectness scores. Following this, Fast RCNN [40] module is utilized to classify the region proposals and to refine the bounding box coordinates. The RPN and the Fast RCNN modules are unified into a single network enabling sharing of convolutional layers (Figure 3.7). The details of Faster RCNN algorithm can be found in [38].

The Faster RCNN algorithm is implemented using TensorFlow open-source library. The input images are horizontally flipped randomly with a probability of 0.5 to execute data augmentation. Subsequently, features are extracted from the input image using a sequence of convolutional layers which are a part of the Inception-ResNet-v2 network. A 3×3 sliding

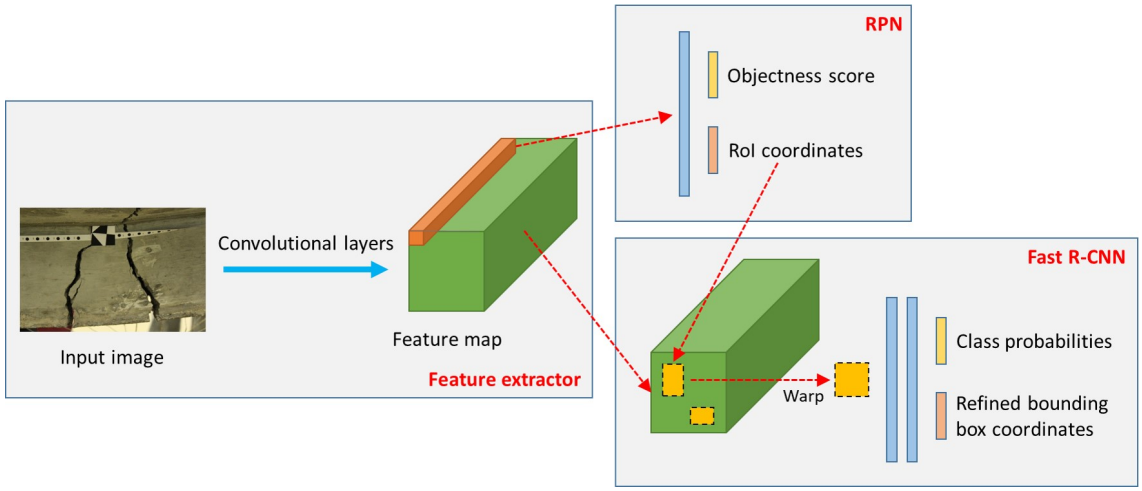


Figure 3.7. Faster RCNN architecture

window is applied to the feature map generated by the last shared convolutional layer mapping it down to a lower dimension. At each sliding window location, a set of 9 anchor boxes having different scales and aspect ratios are considered as region proposals. The anchor boxes are strided by 8 pixels along the height and the width. A large number of region proposals are generated for each image leading to multiple detections. Duplicate boxes are eliminated using a greedy technique called non-maximum suppression (NMS) [48]. The weights of the Inception-ResNet-v2 network are initialized by a model pretrained on MSCOCO data set [49] and fine-tuned thereon using Stochastic Gradient Descent (SGD) algorithm [51] with a momentum value of 0.9. Gradient clipping is employed to avert the problem of exploding gradient. The initial learning rate is set to 0.003 and is gradually reduced thereafter with training steps.

The network is trained on 686 images containing 1023 crack instances. The training data is generated by loading a T-beam as shown in Figure 3.2 and taking pictures of the resulting cracks by means of a movable (hand-held) SLR camera and a camera mounted on a UAV. The performance of the trained network is evaluated on the test data comprising 100 images and 255 crack instances. The test data is produced by loading another T-beam with slightly different cross-section and reinforcement distribution, and photographing the evolving cracks in a similar manner. The predicted bounding boxes (Figure 3.8) are compared with ground truth boxes and the results are reported in terms of precision and recall. The proposed algorithm produces a precision of 95.5 %, which means that 95.5 % of all predicted boxes classified as crack can be designated as correct detections. On the other hand, the recall value is evaluated as 98.6%, indicating that 98.6% of all annotated cracks are correctly detected. It takes roughly about 0.95 seconds at this stage to process a single image of 5184×3456 resolution using a NVIDIA Titan X (Pascal) GPU. It is important to ensure that the predicted bounding boxes enclose respective damage regions completely. This calls for rigorous training of the detection algorithm with stringent requirement imposed on the predicted boxes vis-à-vis overlap with ground truth boxes.

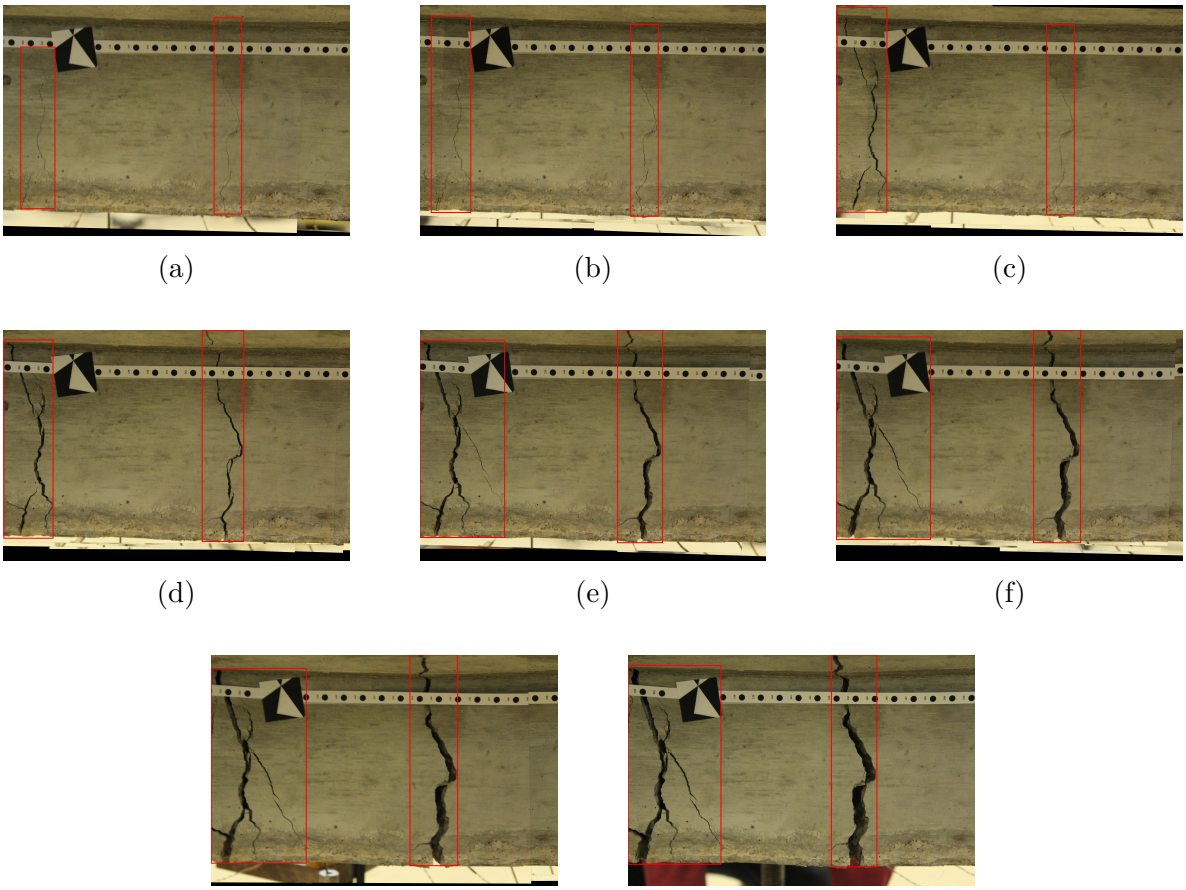


Figure 3.8. Damage chronology produced by successive view synthesis and alignment of correspondences from previous inspection data sets. Cracks detected by Faster RCNN algorithm are highlighted by rectangular bounding boxes.

Damage Segmentation using Morphological Techniques

Morphological approach for image segmentation is motivated by the developments in the fields of set algebra [70] and topology [71]. Dilation and erosion are two rudimentary operations which all morphological methods are based upon. Dilation expands the bright portions of an image, while erosion shrinks the same. These operations applied sequentially form important building blocks for morphological noise removal. Erosion followed by dilation is called morphological opening, which removes bright sharp details from the image. On the other hand, the same operations when applied in the reverse order constitute morphological closing which seeks to remove dark details from an image. Salembier [72] integrated these morphological concepts with bottom-hat transform to propose an algorithm for identification of dark defects in images. The following equation shows a slightly modified version of the algorithm [73] which is used in this study.

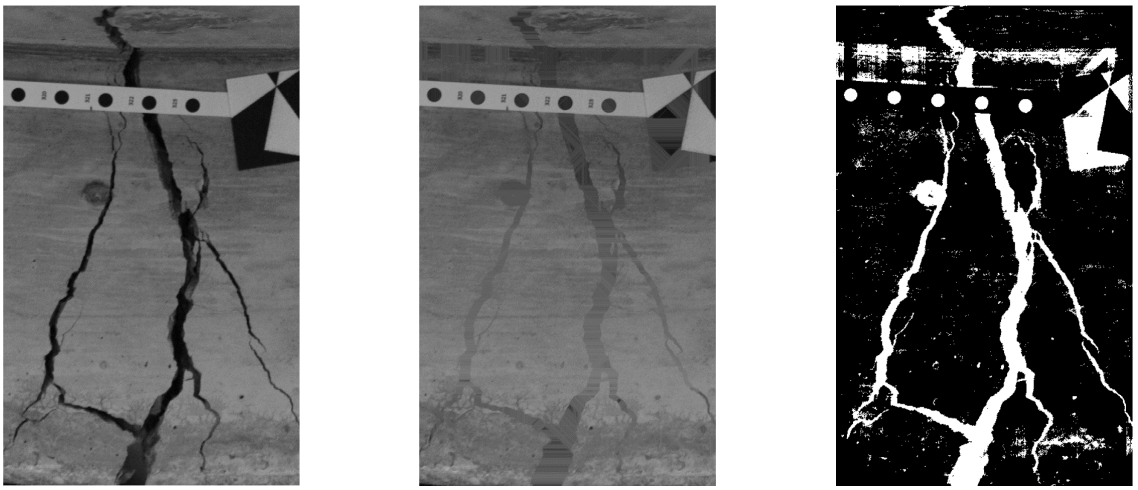
$$T = max[\ (I \circ S_{\{0°,45°,90°,135°\}}) \bullet S_{\{0°,45°,90°,135°\}}, I]\ - I \tag{3.1}$$

where, '$\circ$' and '$\bullet$' denote morphological opening and closing operations, respectively. I is the gray-scale image and S is a structuring element. A structuring element is a matrix that decides which neighbourhood pixels are included in the morphological operations. The structuring element should be suitably chosen as it determines the shape and size of the cracks that can be extracted from an image. Jahanshahi et al. [73] proposed an adaptive approach for estimating an appropriate structural element size based on crack size, camera parameters and camera-to-object distance. A linear structuring element (a structuring element which is line-shaped) with four different orientations ($0°, 45°, 90°, 135°$) is used in this study to make the filter invariant to crack orientation. The crack map so generated (T) was subjected to Otsu's thresholding [74], followed by a series of post-processing and noise removal strategies (Figure 3.9) to obtain the final binary segmentation mask as shown in Figure 3.10. The post-processing scheme involves removing small areas, filling small holes, bridging unconnected pixels, and removing spur pixels and isolated pixels. The entire crack region is segmented in this approach, unlike edge detection-based techniques where only the

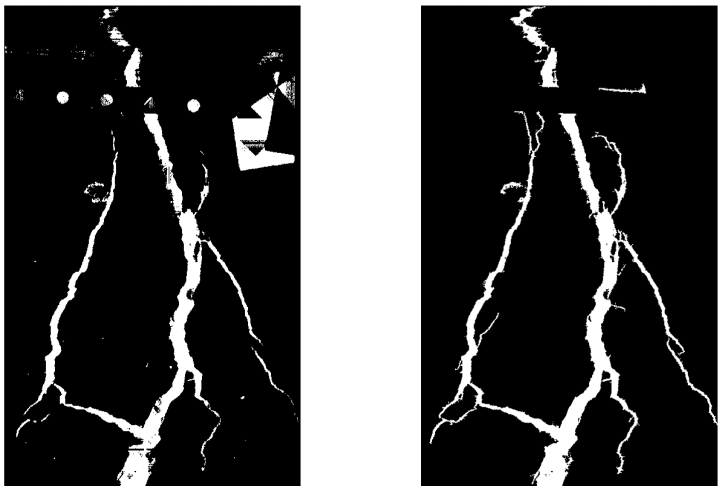
crack boundaries are extracted. Apart from that, morphological approaches are divested of the time-consuming and tedious data annotation and training processes which are required by typical deep learning-based semantic segmentation algorithms.

3.2.4 Damage Quantification

Thickness is an effective indicator for severity of cracks. This section presents the crack thickness quantification algorithm that is used in this study. Some researchers in the past [75] resorted to boundary-to-boundary approach to visually measure the crack thickness. In this approach, the crack thickness at a boundary point is evaluated as the distance to the nearest point on the other boundary. However, the limitation of this approach is that the crack thickness line is usually not normal to the centerline. Moreover, the thickness measured at a boundary point may not be identical to the same measured at the corresponding thickness point located on the other boundary of the crack. These limitations can be redressed by employing centerline-based techniques such as orthogonal line method [8], [73] and distance transform method [59], [60]. The latter approach is adopted in this study. This method begins by finding the centerline of the crack. Researchers in the past exploited various methods for locating the crack centerline in an image. Jahanshahi et al. [10] used fast marching algorithm which was originally proposed by Uitert and Bitter [76]. A number of studies [8], [73], on the other hand, employed morphological thinning operation on binary crack maps, which is followed in this study. The thickness at a given centerline point is given by twice the shortest distance to any of the boundaries. Quasi-Euclidean distance transform is used in this study to find the closest pixel on the boundaries. The effectiveness of the segmentation and quantification approach adopted in this study was previously established by Jahanshahi et al. [73] and the same is not repeated here. Each of the black circles on the tape as observed in Figure 3.8 had a diameter of 5 mm and was represented by 92 pixels in the image. This information is exploited in this study to convert the unit of crack thickness from pixels to mm.



(a) (b) (c)



(d) (e)

Figure 3.9. Steps involved in the segmentation process - (a) Grayscale image (I), (b) Result of $max[\ (I \circ S_{\{0°,45°,90°,135°\}}) \bullet S_{\{0°,45°,90°,135°\}}, I]$, (c) Crack map (T) generated by Equation 3.1, (d) Binary image obtained by applying Otsu's threshold to T, and, (e) Final segmentation mask obtained after post-processing a noise removal.

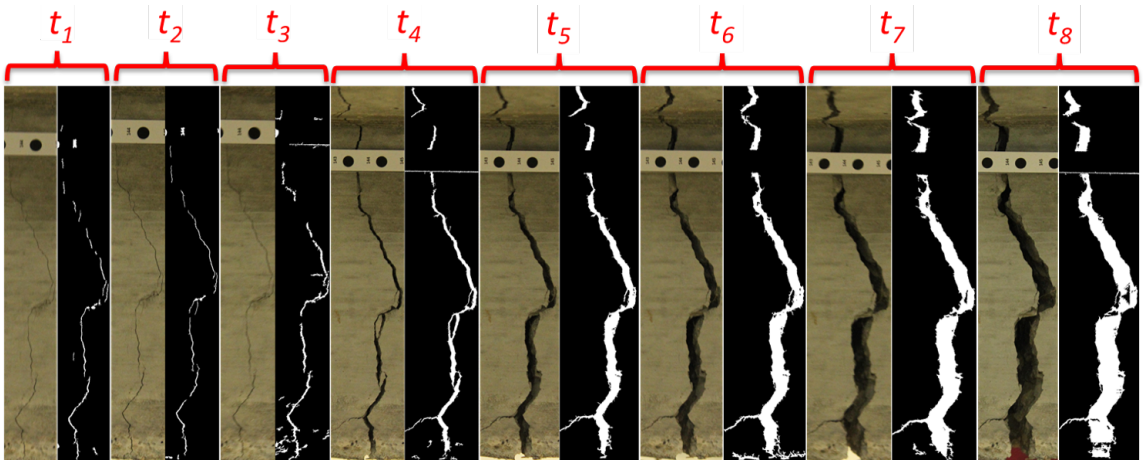


Figure 3.10. Illustrative examples of original image and generated segmentation mask for a crack at different points in time.

3.3 Results and Discussions

The damage evolution dynamics for five different cracks scattered along the span of the tested beam was studied and is shown in Figure 3.11. The distribution of crack thickness evaluated at several points along the centerline of the cracks is plotted against time which is characterized by inspection round. The rectangular boxes denote the range between the first and the third quartiles. The horizontal lines inside the boxes represent the second quartile, also known as the median. The small solid squares inside the boxes symbolize the mean values whereas the whiskers protruding out from the boxes signify one standard deviation on either side of the mean value. All the parameters discussed above are indicators of damage severity and its evolution with time. However, the one parameter which is of highest interest to the inspectors is the maximum thickness. It is represented by small triangles, which are connected by straight lines for better depiction of its evolution with time. It was observed that the maximum as well as the mean crack thickness increases almost monotonically with increase in load. The segmentation algorithm used in this study presumes that the cracks are darker compared to the background. However, this hypothesis is violated at times when light penetrates inside thick cracks making a portion of the crack interior appear bright. This leads to inaccurate segmentation and therefore underestimation of crack thickness as indicated by abrupt dip in the maximum thickness value (Figure 3.11b). However, similar dip observed at lower stages of loading (Figure 3.11a) can be attributed to the debilitating effect of image noise on segmentation of very thin cracks. Increase in load also resulted in higher dispersal in the thickness values due to increase in crack thickness as well as appearance of new branches. Besides, increase in loading intensity increased the difference between the maximum and mean thicknesses.

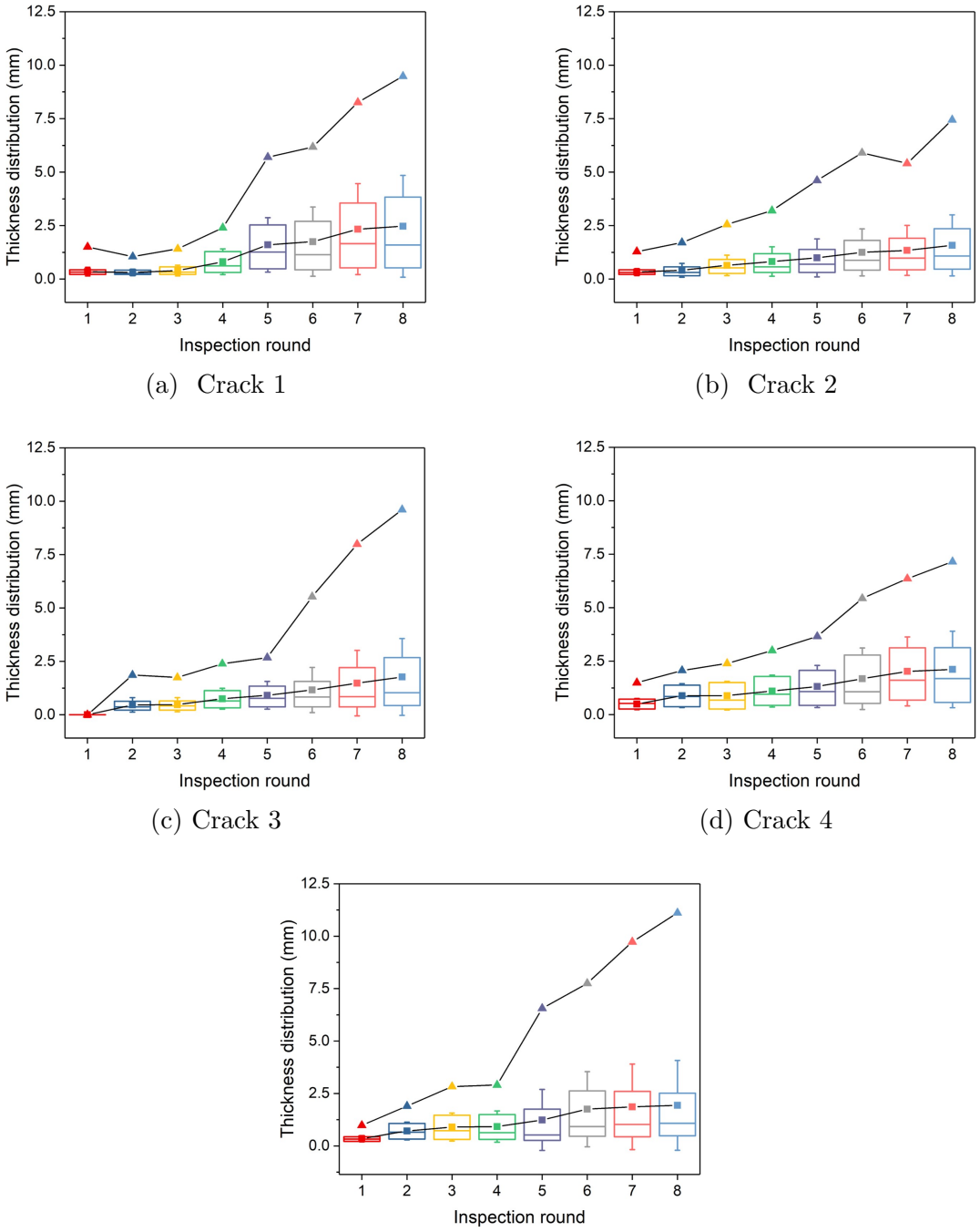
Many a time, evolution of old crack is accompanied by appearance of new branches, which are not accounted for by the maximum thickness. Therefore, total crack thickness together with total area of the cracks, which take into account the main crack as well as its branches, are plotted against time (characterized by inspection round) in Figure 3.12. Total crack thickness is estimated as the summation of crack thicknesses at different locations. On the other hand, the area of a crack is measured by evaluating the number of pixels in a

region enveloped by an 8-connected component in the binary crack map. This figure (Figure 3.12) presents an overall estimate of how fast the crack is growing as a whole.

The deterioration rate (rate of change in crack thickness) is plotted against time (inspection round) in Figure 3.13. It is nothing but the first derivative of crack thickness with respect to time. It is evident from the figure that rate of change in thickness is not monotonic, in contrast with thickness itself. This indicates that the growth rate is not proportional to the applied load. An illustration of this sort will make it possible to single out the two inspection rounds in between which a crack has grown at the fastest pace. For instance, it can be inferred from Figure 3.13 that the crack-5 suffered the worst degradation in between the fourth and the fifth rounds of inspection. Similar conclusions can likewise be drawn for other cracks as well. Such information may prove to be crucial for chronologically connecting the extent of degradation with extreme events from the past such as seismic vibration, fire, mechanical overload, etc. This will facilitate zeroing in on the most probable reason for damage among several possibilities which are otherwise equally likely.

There are occasions when the inspectors are privy to the data recorded by accelerometers or displacement sensors installed in different floors of a building or at different places along the span of a bridge, in addition to images captured by visual sensors. This provides a scope for correlating component level damage severity with peak acceleration or displacement experienced by the structure. Figure 3.14, which shows the variation of crack thickness with the displacement induced by the actuators at loading points, illustrates this concept. The abscissa in this figure should be suitably chosen so as to serve the specific need of the problem at hand. Peak seismic ground motion, mid-span deflection of a bridge or top storey deflection of a building are some of the possible alternatives, to name a few. An analysis as such will enable the structural engineers to anticipate the possible damage in a structure that may be induced by a future earthquake of any given intensity.

All the figures presented in this section provide a clear picture of how fast the crack is growing and thereby facilitate an informed decision making with regard to any immediate follow-up action where necessary. The state-of-the-art approaches for autonomous condition assessment of civil infrastructures are deprived of this crucial time dimension which prohibits any rationale prognostication about an imminent structural failure. However, inclusion of



(a) Crack 1

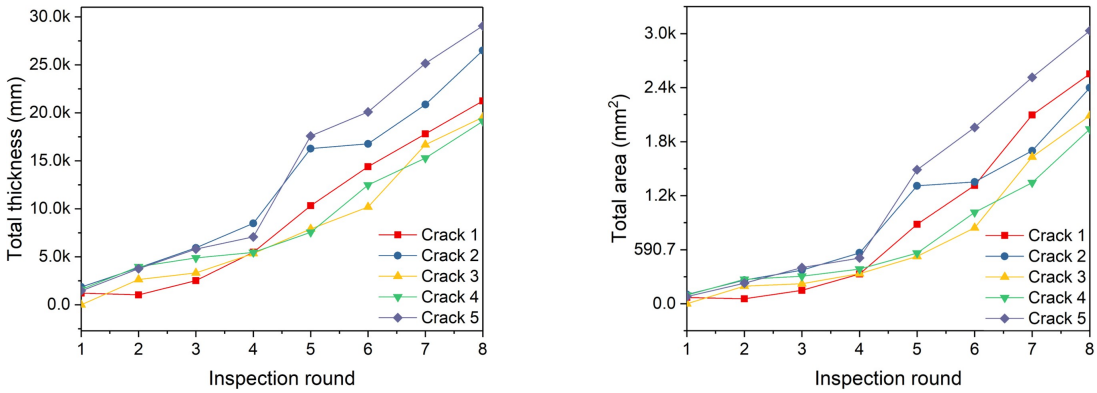
(b) Crack 2

(c) Crack 3

(d) Crack 4

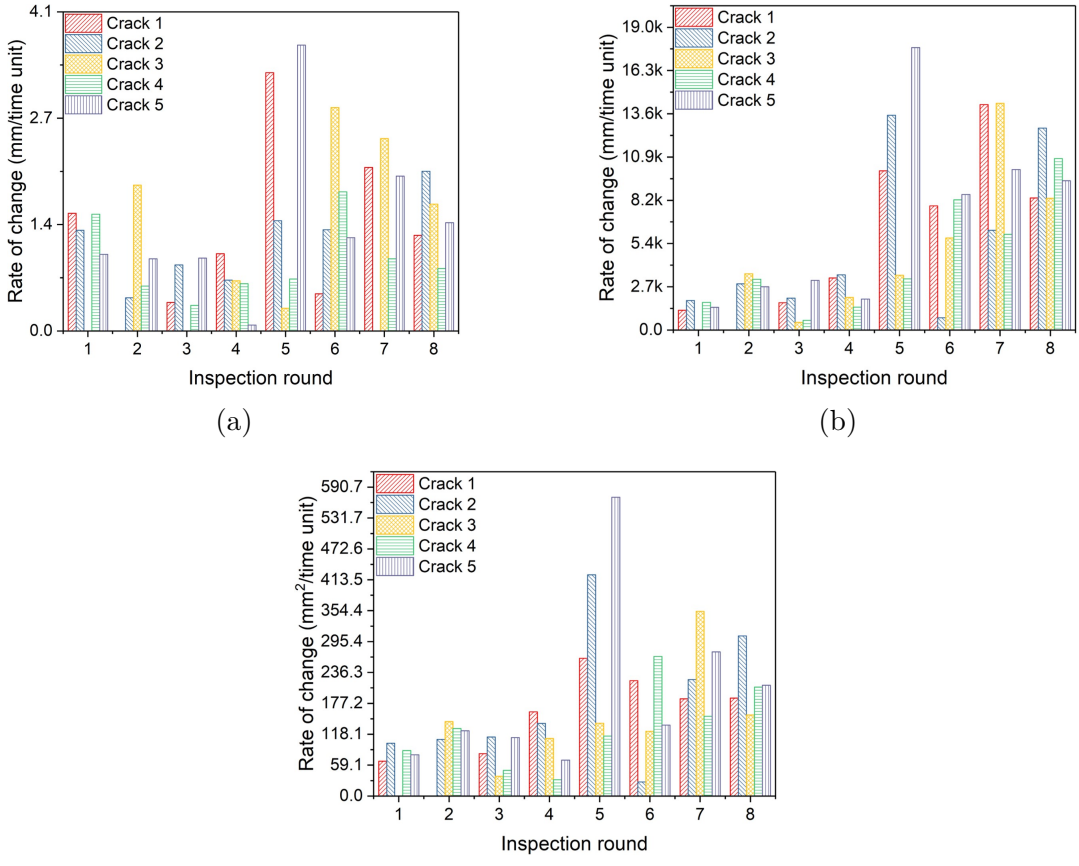
(e) Crack 5

Figure 3.11. Time evolution of crack thickness distribution for five different cracks. The rectangular boxes denote the range between the first and the third quartiles. The horizontal lines inside the boxes represent the second quartile, also known as the median. The small solid squares inside the boxes symbolize the mean values whereas the whiskers protruding out from the boxes signify one standard deviation on either side of the mean value. The small triangles outside the rectangular boxes represent the maximum values.



(a) (b)

Figure 3.12. Time evolution of total crack thickness and area - (a) Total thickness, and, (b) Total area.



(a) (b)

(c)

Figure 3.13. Rate of change of (a) maximum crack thickness, (b) total crack thickness, and (c) total crack area during the intervening time between successive inspections.

precious chronological intelligence, as suggested in this study, into the condition assessment and monitoring framework, will significantly narrow down this limitation, making it possible to estimate the residual life of a structural component and take preemptive measures as needed. It will also be instrumental in recommending any requisite adjustment in the frequency of future routine inspections.

Civil infrastructures undergo visual changes with time due to accumulation of dirt, rust, stains, etc. Even though SURF algorithm is used for feature detection which is invariant to illumination changes, however, the performance of the correspondence identification process can be affected by contamination of visual features leading to reduction in the number of matched points. In extreme cases, this may result in the failure of feature matching and image alignment exercises if there are not sufficient features to estimate the homography matrix accurately. However, in presence of adequate interest points, as in the case of present study, the proposed algorithm will perform reasonably well without any appreciable loss of accuracy. The detection algorithm can be made robust against such surface irregularities by diversifying the training data with regard to all possible noise intrusion and illumination conditions. A lot of noises will be disposed of at this stage by rejection of nonessential background. Finally, proper execution of post-processing strategies in the segmentation module will in effect make the images noise free.

The beam specimens considered in this study were subjected to flexural failure. Therefore, most of the cracks that appeared on the surface were predominantly vertical. However, inspectors often run into situations where structural elements fail in shear giving rise to cracks that are primarily diagonal. In such cases, the Faster RCNN algorithm will predict a larger bounding box enhancing the scope for noise ingression. However, an appropriate post-processing in the segmentation stage will ensure that the noises are duly identified and eliminated. Moreover, linear structuring element with different orientations (Equation 3.1) renders the segmentation technique invariant to crack orientation. It is not uncommon to encounter situations where two initially unconnected cracks intersect and become inseparable with increase in load. In such situations, it is recommended that the pair of cracks should be treated as a single entity and evaluated jointly. Cracking in concrete is used as a case study to validate the efficacy of the proposed approach. However, the same tech-

niques can be extended to any other defect category or to multiple defect categories with appropriate modifications. The data sets used for training and validation of the detection algorithm should be suitably updated to include instances from all the defects being investigated. Defect-specific segmentation and quantification algorithms should be invoked to put in place a comprehensive condition assessment pipeline.

3.4 Conclusion

This study was motivated by the observation that most of the published works in the area of vision-based autonomous structural inspection and health monitoring are agnostic to the time dimension. Ignoring vital historical information, which can otherwise be a key to time-based analysis of damage growth, makes it impossible for inspectors to act preemptively to avert any imminent structural failure and consequent human and financial losses. This study aimed at filling this research gap by proposing a novel computer vision-based approach to leverage from the crucial chronological intelligence embedded in archival images captured by mobile inspection robots or UAVs. Strategies are proposed for autonomous exploration into the erstwhile inspection data looking for correspondences, view synthesis from multiple correspondences and alignment to the current scene under consideration, localizing damage in the reconstructed scenes from the past, segmenting damage, and finally quantifying the damage to extract necessary information and derive meaningful conclusions, after a damage is detected in the current data set. Time history of damage is graphically presented facilitating easier interpretation in addition to predictive and quantitative evaluations. The proposed framework will also enable a transition from the current schedule-based inspection process, where a structure is inspected at regular time interval, to a condition-based inspection paradigm where the frequency of inspection can be adjusted based on the current state of an infrastructure system. Cracks on concrete surface are used as a case study to demonstrate the feasibility of this approach, which can be potentially extended to any type of structural defects, namely, spalling and corrosion (Appendix A). However, effective implementation of the proposed algorithm makes it necessary to have complete coverage of the damaged areas and adequate overlap between successive images at each batch of inspection

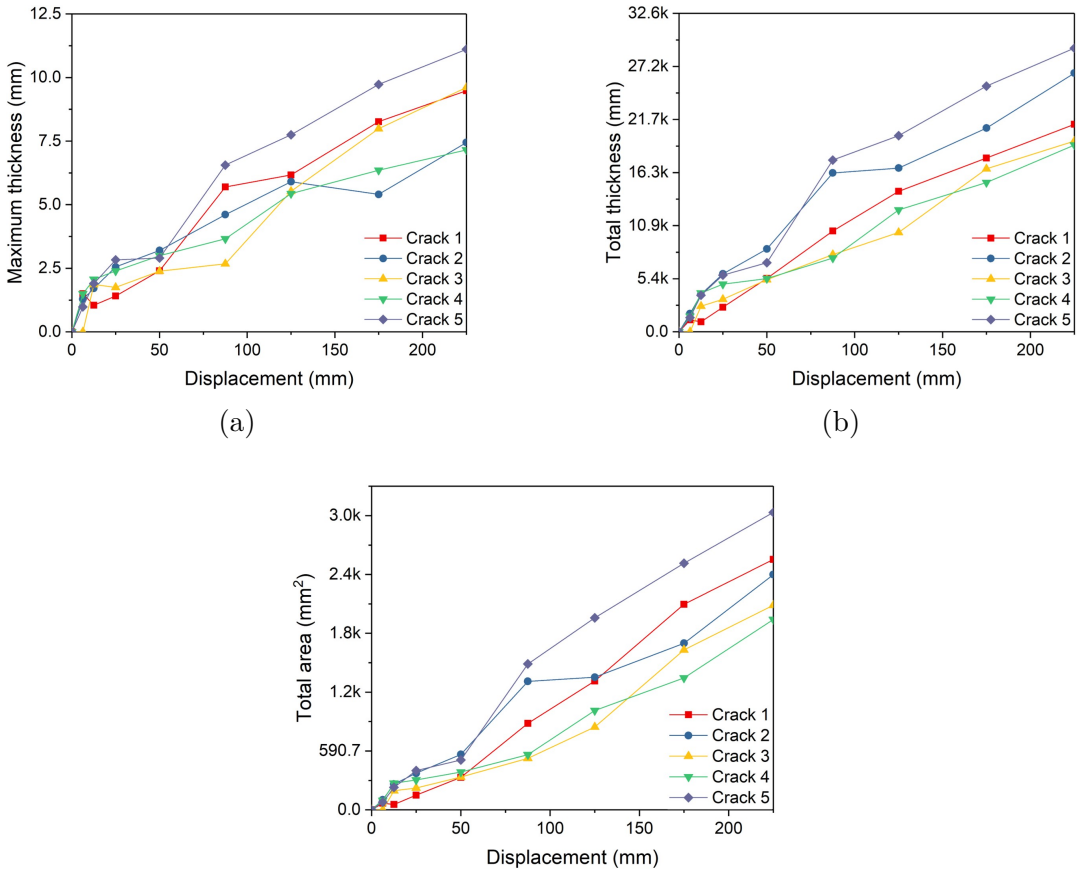


Figure 3.14. Variation of (a) maximum thickness, (b) total thickness, and (c) total area of cracks with respect to induced displacement.

data. Besides, this algorithm will fail if there are insufficient visual features or if the structure under investigation is not planer. Incorporation of IMU and GPS information may lead to more robust scene reconstruction in such situations, which is a scope for further research.

4. MULTI-SENSOR FUSION FOR DEEP LEARNING-BASED AUTONOMOUS DAMAGE DIAGNOSIS EXPLOITING SYNTHETIC TRAINING DATA

4.1 Background

Convolutional neural networks (CNN) are widely used by the computer vision community for object recognition and semantic segmentation from visual data. The state-of-the-art algorithms in this area largely rely on photometric measurement of colors due to immediate availability of inexpensive high resolution cameras. However, regular cameras translate 3D scenes into 2D spaces leading to a loss of information vis-à-vis distance and scale. This imposes various limitations which come in the way of realizing the full potential of vision-based techniques. Therefore, arrival of commercially available low-cost depth sensors such as Microsoft Kinect was believed to unlock the next level of computer vision applications which propelled the scientific community to explore various utilities of depth measurement. A number of investigations to this end revealed that integration of depth information significantly enhances the performance of state-of-the-art detection and segmentation models. Schwarz et al. [77] demonstrated that infusion of depth data improves the efficiency of CNN-based robotic scene understanding and manipulation in disaster response scenario. Hazirbas et al. [78] proposed an encoder-decoder architecture called FuseNet for semantic segmentation of indoor scenes and observed that fusion of depth information significantly augments the segmentation accuracy. Similar observations are recorded by Li et al. [79] who invoked long short-term memory (LSTM) netowrk for multi-modal indoor scene labelling. Wang et al. [80] proposed a novel feature transformation network for fusion of multi-modal sensor data. It connects a set of convolution and deconvolution layers which have the same structure for different modalities. The feature transformation network extracts common features which are shared between different modalities as well as modality specific features which capture visual patterns visible only in one modality. Park et al. [81] demonstrated that fusion of depth features enhances the accuracy of indoor semantic segmentation by proposing a multi-level feature fusion network called RDFNet taking advantage of residual learning

and skip connections. Locally sensitive deconvolution network was exploited by Cheng et al. [82] for semantic segmentation of indoor scenes. The authors introduced gated fusion layer which effectively combines features from various sensor modalities. The proposed approach increased the segmentation accuracy significantly. Xu et al. [83] adopted shared weights strategy and parameter-free correlation of modality-correlated and modality-specific features for multi-modal object detection. The authors noticed an overall improvement in detection accuracy owing to multi-sensor fusion. Ophoff et al. [84] exploited single-pass CNN architecture to fuse depth and visual sensor data for real-time pedestrian detection leading to an improvement in accuracy. It may be noted here that contrary to standard practice of boosting model accuracy by increasing model complexity, depth fusion seeks to provide a less expensive alternative for enhancing model performance by enriching the information content of input data. Other notable works in this area include [85]–[88].

Aging civil infrastructures require periodic inspection in order to prevent sudden failure which causes loss of lives and economic setbacks. The existing inspection techniques are by and large manual and therefore time consuming, subjective and risky. Computer vision-based algorithms have been explored in recent times to investigate the prospect of robotic inspection as a viable alternative to such manual techniques. A number of studies ([5]–[7], [30], [33], [69], [89]) exploited deep learning (DL)-based methods to this end for autonomous defect detection in civil infrastructures. However, the previous studies relied solely on the photometric data for identifying damages in videos and images. Only a few studies in the past focused on investigating the effect of depth fusion on the performance of RGB-based DL algorithms vis-à-vis damage detection in civil infrastructure systems. Zhou and Song [90] probed into the fusion of intensity and range images to train a CNN for classification of roadway cracks. In separate studies, the authors [91], [92] relied on fully convolutional encoder-decoder networks for semantic segmentation of concrete roadway cracks based on range images alone. A number of studies ([73], [93], [94]), on the other hand, exploited depth measurement for quantifying different types of defects such as concrete cracks, spalling, etc. However, there is no study till date which leveraged depth perception for semantic segmentation of various damages that commonly occur in RC structures subjected to extreme loading. Depth information may prove to be crucial for distinguishing between actual damage and

damage-like artifacts having similar visual features. It can also serve as a force multiplier when it comes to damage detection under poor lighting condition. Jahanshahi et al. [95] proposed a pavement defect segmentation technique by fitting a plane to depth values measured by inexpensive depth sensors and detecting the defective area by thresholding the relative depth evaluated with respect to the fitted plane. This unsupervised approach is fraught with many limitations and may fail in situations where the inspection surface is not planar. Moreover, it is not endowed with the capability of distinguishing between multifarious damage categories. This leaves an information gap in the existing knowledge base which the present study aims to fill. This study incorporates depth fusion into a fully convolutional network (FCN) for semantic labelling of different damage types relevant to RC structures. Fusion of depth data was observed to enhance the segmentation performance significantly.

The single biggest factor that deterred the scientific community from exploring the utility of depth data with regard to vision-based autonomous condition assessment of civil infrastructure is the scarcity of a publicly available damage data set that contains depth information. This shortcoming is overcome in this study by using state-of-the-art computer graphics techniques [96] to generate a database of synthetic damage containing spatially aligned RGB and depth information following guidelines from ACI 318 [97]. Recent studies [98] have shown that DL algorithms trained on synthetic data perform reasonably well on real data. Synthetic data in the past has been exploited by Hoskere et al. [99], [100] for semantic segmentation of cracks and corrosion in a meter gate, and by Narazaki et al. [101] for recognition of bridge components and damage severity levels in high-speed railway viaducts. The authors of these studies superimposed surface textures of damage on the graphics models of various structural elements to impersonate a photo-realistic appearance of damage. The present study, on the other hand, relied on reconstructed 3D models of real buildings and induced pit-like depressions on the surface akin to the spalling of concrete in RC structures. Three different damage categories are considered in this study for semantic labelling, namely, spalling, spalling with exposed rebars, and severely buckled rebars (Figure 4.2). An extensive discussion on the data generation process is included in Section 4.2.

Representing the depth information in a proper way is paramount for getting the most out of depth fusion. The quest for a suitable strategy for representing depth data has led

to the emergence of various encoding techniques such as Absolute Depth-based Encoding (ADE) and Surface Normal-based Encoding (SNE). Apart from these two existing encoding techniques, this study also proposes a novel two-stage approach based on depth values measured relative to a fitted plane. This encoding technique is referred to as Relative Depth-based Encoding (RDE) in the current study. The central ideas behind all three encoding techniques are elaborated in Section 4.3. A comparative performance evaluation revealed that SNE outperforms other two encoding techniques in terms of accuracy and robustness.

Encrypted depth (ED) is a general term used in this study to describe absolute depth (AD), surface normal (SN), and relative depth (RD) for respective encoding techniques. Stacking the RGB and ED images right at the beginning and processing the resulting multi-channel image is a time-honored fusion strategy. However, previous studies indicate that an early fusion as such does not always lead to the optimum performance. This study therefore explores various other fusion strategies as described in Section 4.4. It is observed that the best fusion strategy is dependent on the depth encoding technique. Several experiments are conducted in this study for comparative assessment of different depth encoding techniques and fusion strategies. The results are presented and discussed at length in Section 4.5.

Despite proven advantages of depth fusion, it is not to be forgotten that depth sensors are not yet as pervasive and ubiquitous as RGB cameras. Moreover, practical application of depth sensing during real robotic inspection has many challenges. The traditional lidar-based depth sensors are generally large and weighty, and therefore not suitable to be integrated with mobile robotic platforms. The recent consumer-grade depth sensors, on the other hand, have the advantages of being lightweight and low-cost. However, many of these sensors exploit laser scanning techniques which are susceptible to interference by sunlight, leading to a poor outdoor performance. Besides, depth sensors may lead to increased energy consumption reducing the operating life of UAVs, which rely on on-board batteries as primary energy sources. This reduces the efficiency of robotic inspection by enhancing the inspection time and costs. In view of these practical constraints, it will be ideal to forego depth sensing at test time without foregoing the benefits of depth fusion. This study will aim to achieve this research objective by leveraging two important advances in the area of multi-modal sensing, namely modality hallucination [102], [103] and single view depth estimation [104],

[105]. These surrogate techniques enable us to simulate depth maps or hallucinate mid-level convolutional depth features from single-frame RGB images. This eliminates the need of depth sensing at test time without considerably undermining the segmentation performance, as illustrated in Section 4.6. In the end, the conclusions of this study are summarized in Section 4.7.

4.2 Generation of Synthetic Data

4.2.1 Basic theory

Houdini [106] is a 3D animation and visual effects software developed by SideFx which is used in this study for the generation of the synthetic data. What makes Houdini distinct from other commercially available computer graphics software is its procedural generation capability [107] through Procedural Dependency Graph (PDG). It enables parallelization of sequential workflows enabling automation and scaling of the entire pipeline. Complex dependencies are described visually with nodes and transformed into a set of actionable, schedulable tasks, which are then distributed with the help of a scheduler and computed in parallel. Synthetic damages are induced by Boolean subtraction of a solid geometry from baseline objects. The baseline objects (Figure 4.1(a)) considered in this study are 3D reconstructions of real buildings obtained from Matterport3D dataset [108]. Matterport3D is a large-scale RGB-D dataset that comprises 10800 panoramic views from 194400 RGB-D images of 90 building-scale scenes with precise global alignment. The geometry that is subtracted from the baseline structure can be something as simple as a solid sphere (Figure 4.1(b)). However, to create an appearance of real damage, Simplex noise is applied to the sphere (Figure 4.1(e)) prior to Boolean subtraction. Simplex noise is a procedural texture primitive very similar to Perlin noise [109] but with fewer directional artifacts and lower computational overhead in higher dimensions. The shape and size of the sphere and various noise parameters can be randomized for different iterations. A region of interest, where a damage is intended to be induced, is manually demarcated and dissociated from the remaining model (Figure 4.1(d)). The solid geometry is then placed on the isolated wall section so as to ensure that they intersect each other (Figure 4.1(g)). The portion of

the wall intercepted by the solid geometry is then cut out (Boolean subtraction) leading to a configuration shown in Figure 4.1(h). This modified wall section is then merged with the rest of the structure to retrieve the entire building model (Figure 4.1(k)). To assign photo-realistic material using principled shader, a real image of concrete spalling (Figure 4.1(j)) is projected on the damaged part of the building model as shown in Figure 4.1(l). The reinforcement bars (Figure 4.1(f)) are modelled using polywire nodes which are used to create complex tubular geometries with smooth bends and intersections. The reinforcement bars are textured (Figure 4.1(i)) and placed on the damaged part of the building resulting in a realistic scene of concrete spalling with exposed rebars. As a whole, 629 of such scenes are generated containing complex realistic backgrounds and three different damage categories, namely, spalling, exposed rebars, and severely buckled rebars (Figure 4.2). Each scene is rendered from different camera positions and orientations using a ray tracing [110] or physically based rendering [111] engine resulting in 1789 pairs of RGB and depth images synchronized in space and time. A histogram of camera-to-damage distance for all such images is presented in Figure 4.3. The figure indicates that most of the images are rendered from a distance of 4-5 meters, which is consistent with real robotic inspections.

4.2.2 Automatic Labelling

DL-based segmentation models require manual labelling of the training and test data which is an arduous task. Also, any inaccuracy in the manual labelling adversely impact the segmentation performance. However, synthetic data generated by computer graphics software can be labelled automatically which enhances the speed and accuracy of the annotation process. Houdini has a special feature which assigns a unique identification number to each material in a model. This enables the pixel coverage for each material to be readily available based on this identification number. The damaged regions in each scene are identified in this manner and labelled according to the severity of the damage.

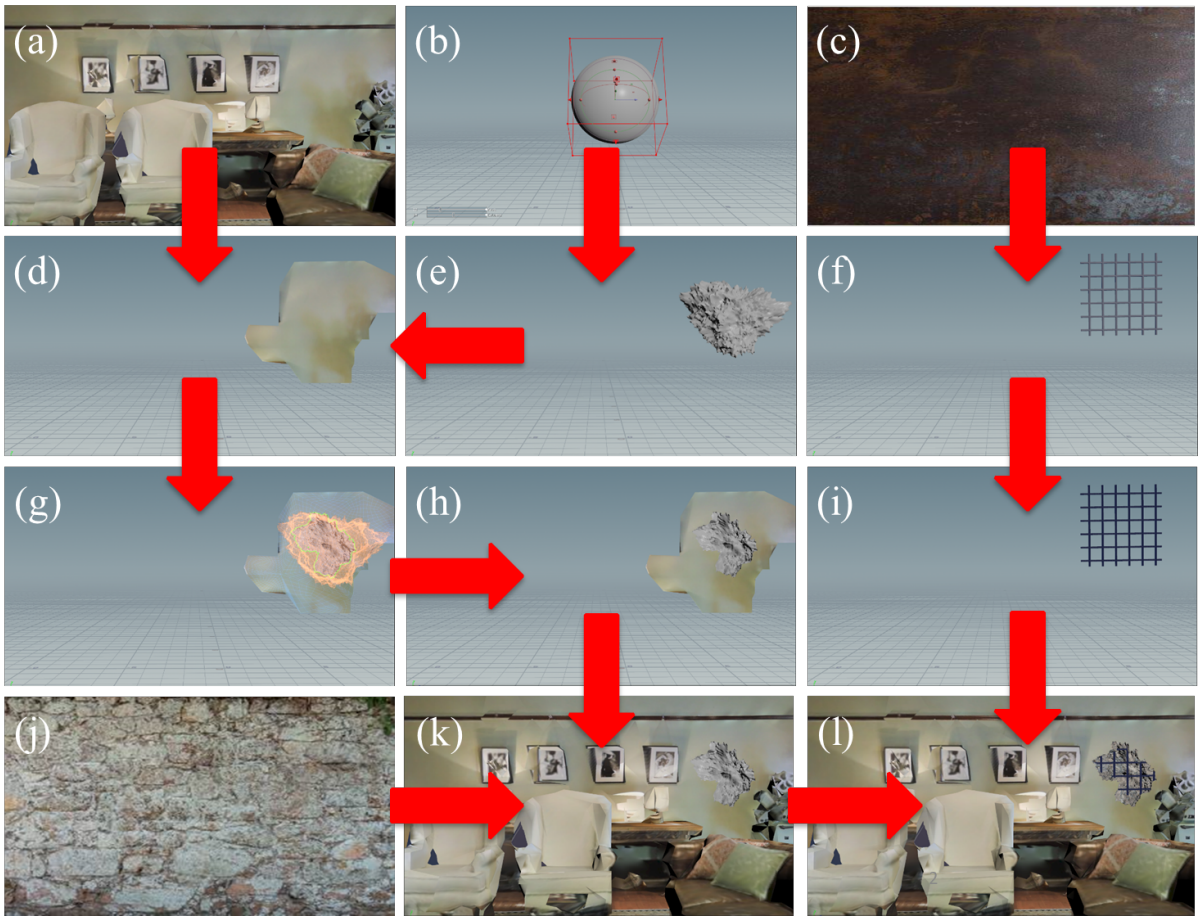


Figure 4.1. Synthetic damage data generation pipeline using computer graphics tool

(a)

(b)

(c)

Figure 4.2. Damage categories considered in this study - (a) Spalling, (b) Spalling with exposed rebars, (c) Severely buckled rebars.

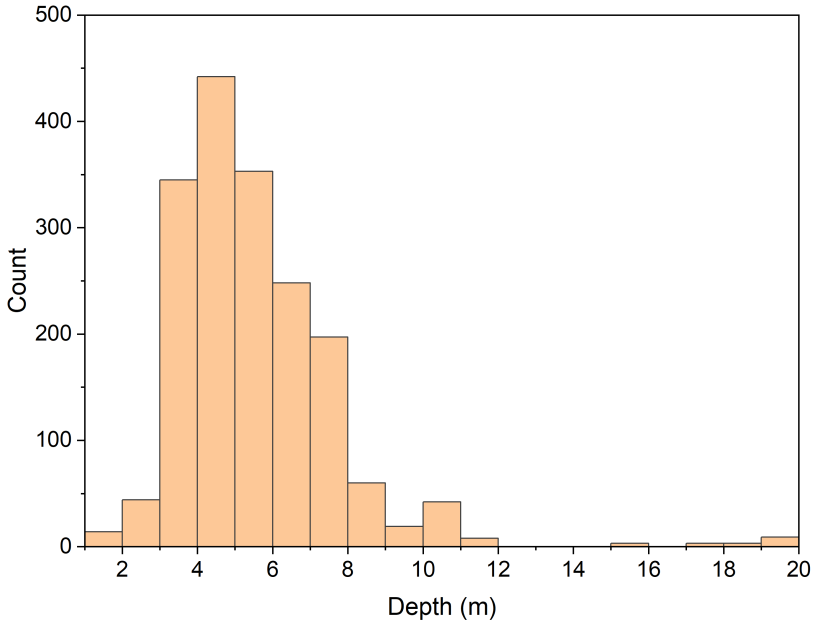
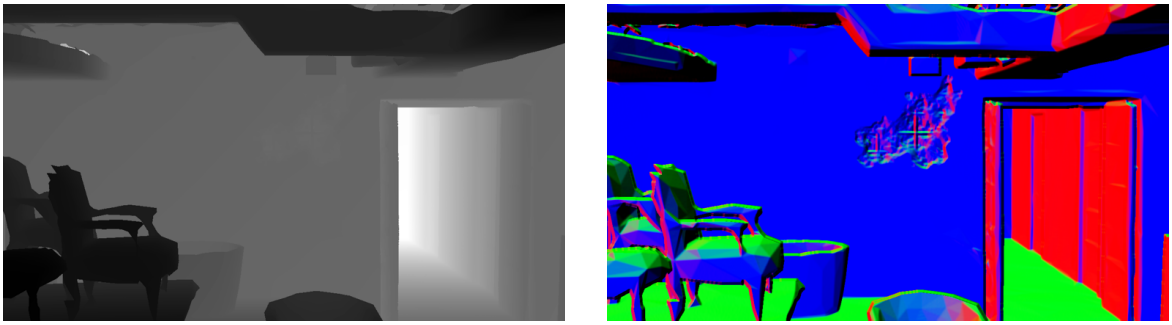


Figure 4.3. Histogram of the distance between the camera to the center of a damage

4.3 Depth Data Encoding

The value stored at each pixel of a raw depth map (Figure 4.4a) represents the absolute distance between the camera and a physical point in the 3D space, projected on the optical axis of the camera. Conventionally, it is the Z coordinate of a 3D point with respect to a coordinate system having its origin at the camera center and its Z axis aligned with the principle axis of the camera. The brighter pixels in the depth image are away from the camera and the darker pixels are closer to the camera. Normalized depth map has been used by many researchers in the past as input to neural networks needing no preprocessing of data. This encoding technique is referred to as Absolute Depth-based Encoding (ADE) in this study. A number of studies, on the other hand, applied image processing techniques to extract useful features from the depth map, which were then input to neural network models. One of the features that are widely used in this regard is based on SN vectors which reveal the shape of a 3D object. Knowing the focal length of the camera and the depth value at each pixel, the 3D position for each point in the scene can be ascertained using a pinhole camera model. This is followed by fitting of a local plane at each physical point with the help of its neighbouring points. The size of the neighbourhood is generally hand-engineered based on the level of noise in depth measurement. The equation of the fitted plane can be used to deduce the SN vector at each point. A SN map (Figure 4.4b) is thus produced having three channels representing the three components of the SN vector estimated at each pixel of the depth map. This approach for representing depth data is categorized as Surface Normal-based Encoding (SNE) in the present study. The generated SN map looks like a texture, where all points lying on a plane, having the same SN vector, are represented by the same color. However, different points in a damaged region have different SN vectors. Therefore, the uniformity in the texture is lost in this region. This provides a informative cue about the presence of a damage in the scene.

The depth of a damage is usually small compared to the camera to object distance. Therefore, in presence of noisy data, the additional depth due to a damage may not make a mark when an image is captured from a relatively large distance. Therefore, this study proposes a novel two-staged encoding technique based on RD. In this technique, the potential



(a) Absolute depth map (b) Surface normal map

Figure 4.4. Various depth encoding techniques - (a) Absolute depth-based encoding (ADE), (b) Surface normal-based encoding (SNE).

damage areas are first identified using a RGB-based detection model. A region-based CNN popularly known as Faster R-CNN [38] is exploited in this study to this end. In Faster R-CNN, the input image is first passed through a series of convolutional layers. The feature map generated by the last convolutional layers is sent to a region proposal network for generating regions of interest, which are finally refined and classified using a Fast R-CNN [40] module. The classification of interest regions at this stage is binary (presence or absence of a damage) and agnostic to the severity of damage. The geometry of the detected interest regions is then estimated with the help of a pinhole camera model utilizing the depth map and the focal length of the camera. Following this, a plane is fitted to the 3D points which represents the undamaged surface encircling a damaged region. A RANSAC-based [67] plane fitting technique suggested by Jahanshahi et al. [95] is adopted to this end in this study. The distance of all points in the damaged region from the fitted plane represents the relative damage depth as shown in Figure 4.5. However, the use of this technique is limited only to situations where the damage is located on a planar surface such as walls, roofs, etc. It will fail to fit a plane if the damage is located at the corner of a room or at the intersection between two structural members. This encoding technique is referred to as Relative Depth-based Encoding (RDE) in the current study.

4.4 Fusion Strategies

A fully convolutional encoder-decoder network (Figure 4.6) proposed by Hazirbas et al. [78] is used in this study as a baseline model to investigate the effect of depth fusion. The encoder part, which extracts features from the input image, resembles a VGG-16 architecture [112] without the fully connected layers. It is pre-trained on ImageNet dataset [113] and fine-tuned thereon. The decoder part, on the other hand, upsamples the feature maps back to the original input resolution using memorized unpooling [114]. Various fusion techniques are explored in this study to identify the best strategy. A pure RGB-based network, as shown in Figure 4.6a, is used as a benchmark to assess the benefit from depth fusion. The architecture shown in this figure (Figure 4.6a) is also valid for a network trained solely with ED data. Figure 4.6b depicts early fusion (EF), where the RGB and ED images are stacked and the

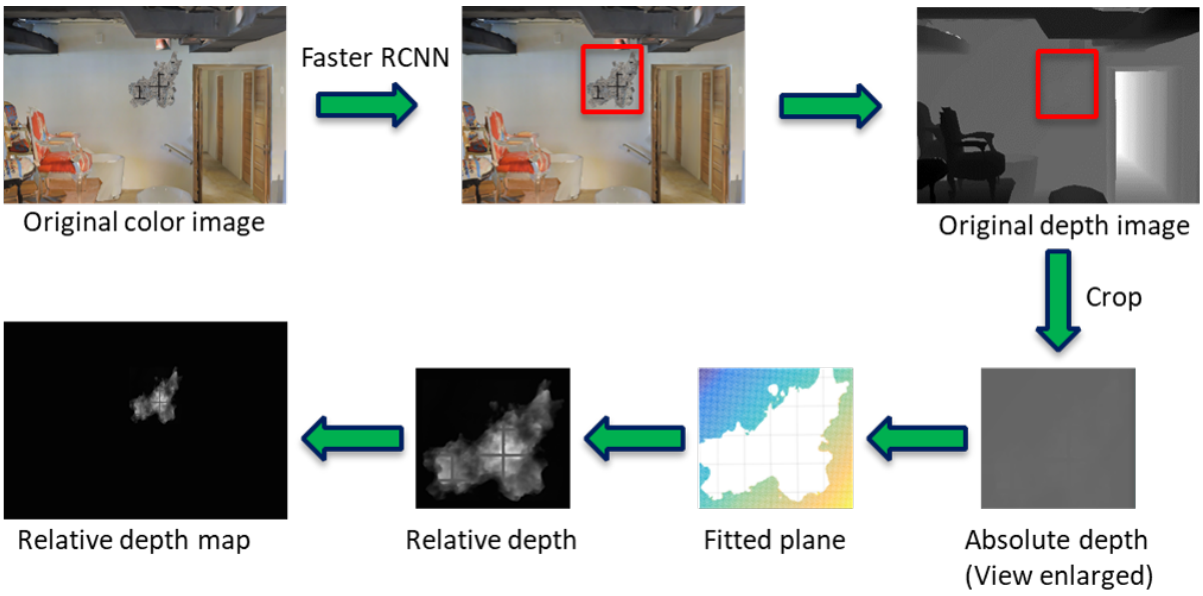


Figure 4.5. Relative depth-based encoding (RDE) - Flowchart depicting the process of relative depth map generation

resulting multi-channel image is sent to the FCN to generate an output segmentation mask. In late fusion (LF) presented in Figure 4.6c, the RGB and ED images are passed through two separate encoder and decoder pairs, and the classification score maps produced by the last decoder layers are added up to produce the final semantic labels. Additionally, a number of intermediate fusion (IF) strategies are investigated which entail separate encoders but a shared decoder for RGB and ED channels. The feature maps from the ED branch of the encoders are fused (element-wise summation) to respective feature maps in the RGB branch at different levels. A collection of 31 different fusion strategies can be put in place by considering all possible combinations of five IF levels. Examining the efficacy of all potential fusion strategies is an arduous task and therefore beyond the scope of the present study. Alternately, a small number of fusion schemes are hand-picked from the larger group of 31 fusion strategies, and are investigated in this study. The selected fusion strategies are denominated by numbers indicating the constituent fusion levels (FL). Figure 4.6d presents an illustrative example of IF strategies where features from the RGB and ED branches are fused at all intermediate levels.

4.5 Results and Discussions

Several experiments are conducted in this study for comparative analysis of different depth encoding techniques and fusion strategies. Five-fold cross validation is conducted to estimate the robustness of various fusion schemes. The size of training and test data used for each cross-validation round is shown in Figure 4.7. It should be noted that there was no overlap between the test data used in different rounds of cross-validation. The depth data generated by computer graphics software are generally noiseless. However, practical depth sensors invariably contain some sensor noise. Therefore, a zero-mean Gaussian noise is added to the actual depth values to mimic a real world depth sensing. The noise was assumed to be a random variable having zero mean and standard deviation varying with the sensor-to-object distance. This variation in standard deviation with sensor-to-object distance was experimentally measured by Zennaro et al. [115] for the first generation Kinect sensor (Kinect v1) as shown in Figure 4.8. A quadratic fit to the observed data is used in this

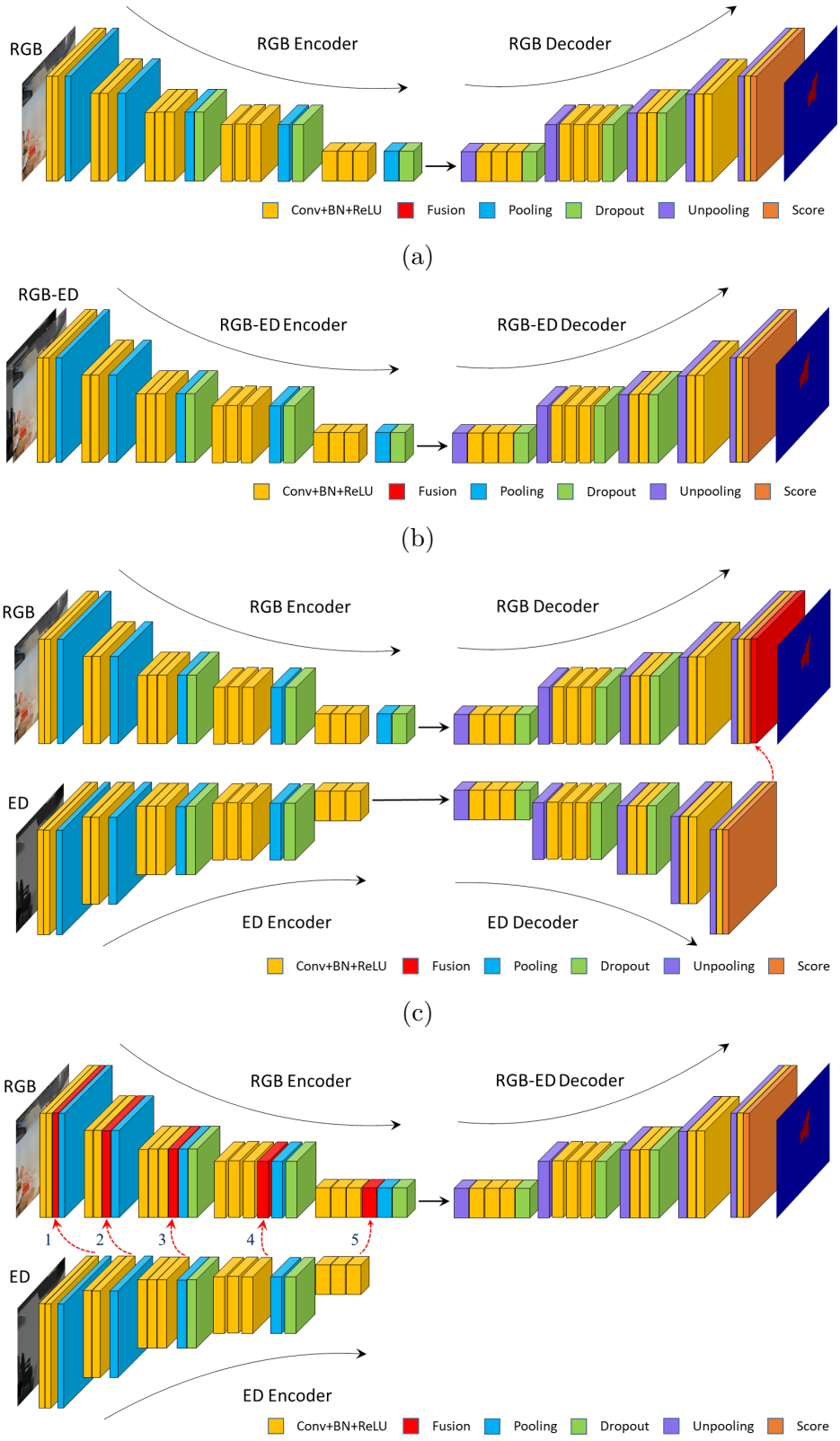


Figure 4.6. Various fusion strategies considered in this study - (a) No fusion, (b) Early fusion (EF), (c) Late fusion (LF), (d) Illustrative example of intermediate fusion (IF).

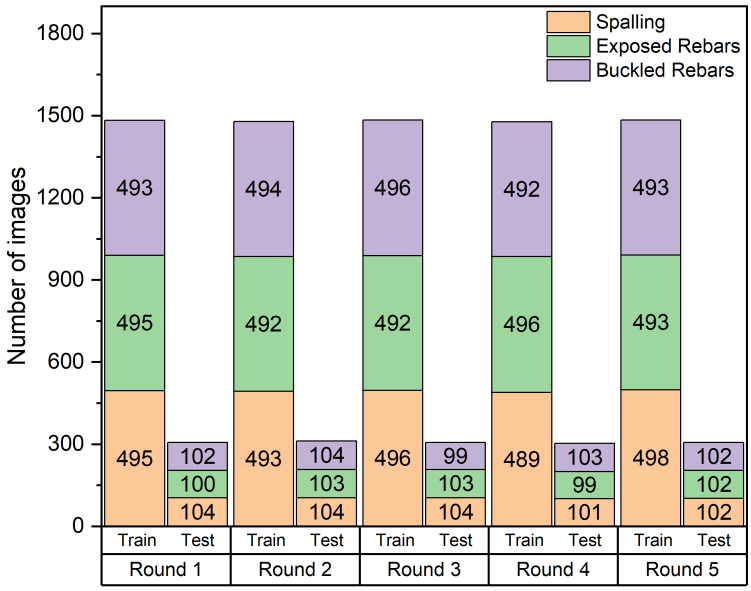


Figure 4.7. Category-wise training and test data size for different cross-validation rounds

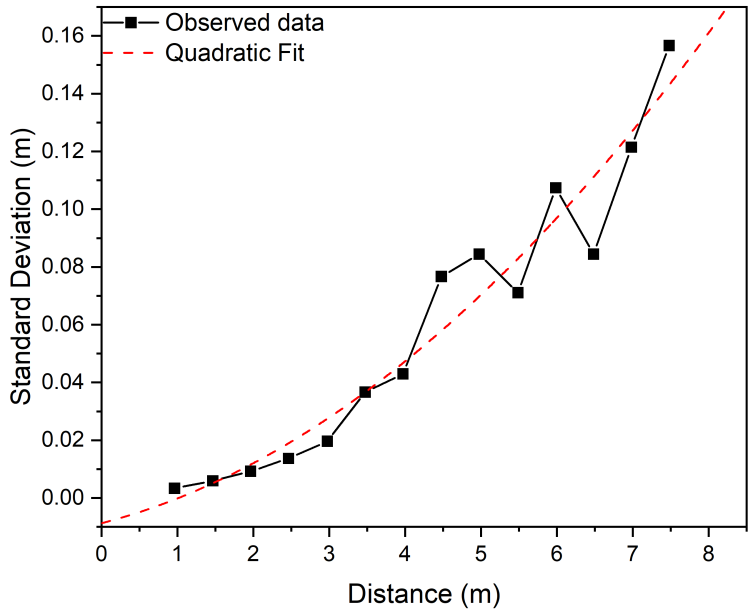


Figure 4.8. Sensor noise in depth measurement by Kinect v1 [115]

study, without a loss of generality, to predict the noise level at a given depth. At each pixel of a depth map, the noise level is computed by random sampling from a zero-mean normal distribution, the standard deviation of which is dictated by the quadratic fit presented in Figure 4.8. The estimated noise is then added to the actual depth value to obtain the noisy AD data. Subsequently, SN maps and RD maps are computed based on this noisy depth data following the procedures described Section 4.3.

The performance of the models is evaluated in terms of intersection over union (IoU) of the predicted and ground truth damage regions. The variation range of overall IoU, which is the average of class-specific IoUs, as obtained from the cross-validation is shown in Figure 4.9. In this figure, the small squares inside the rectangular boxes denote the mean values. The horizontal lines inside the boxes represent the median values. The lower and upper sides of the boxes signify one standard deviation on either side of the mean values. The whiskers protruding out of the boxes indicate the minimum and maximum values of overall IoU. This figure reveals that depth fusion, regardless of the encoding technique, significantly enhances the performance of traditional RGB-based segmentation models, which produced a mean overall IoU of 0.690. The EF exhibited the highest accuracy for ADE producing a mean overall IoU of 0.890, which implies a 20% increase in accuracy compared to the baseline RGB-based model. On the other hand, for SNE, the optimum performance was rendered by IF-3, which produced a mean overall IoU of 0.938, indicating a whopping 25% jump in accuracy. Similarly, in the case of RDE, IF-12 was found to exceed other fusion strategies with a mean overall IoU of 0.909, which is 21% higher than the single-modality RGB-based model. It is worth noting that several fusion schemes considered in this study outperform IF-1234 which was identified as the best fusion strategy by Hazirbas et al. [78]. The values of precision and recall averaged over five rounds of cross-validation for Faster RCNN-based regions of interest detection in RDE were 0.987 and 0.995, respectively.

Training based on AD data alone did not converge and therefore produced no meaningful results. This is because of the inconsequentiality of the additional depth due to a damage as compared to the much larger sensor-to-object distance and the diminution in its discriminating ability in presence of sensor noise. This also triggered non-convergence of the network with LF. This implies that AD can provide complementary information, but cannot

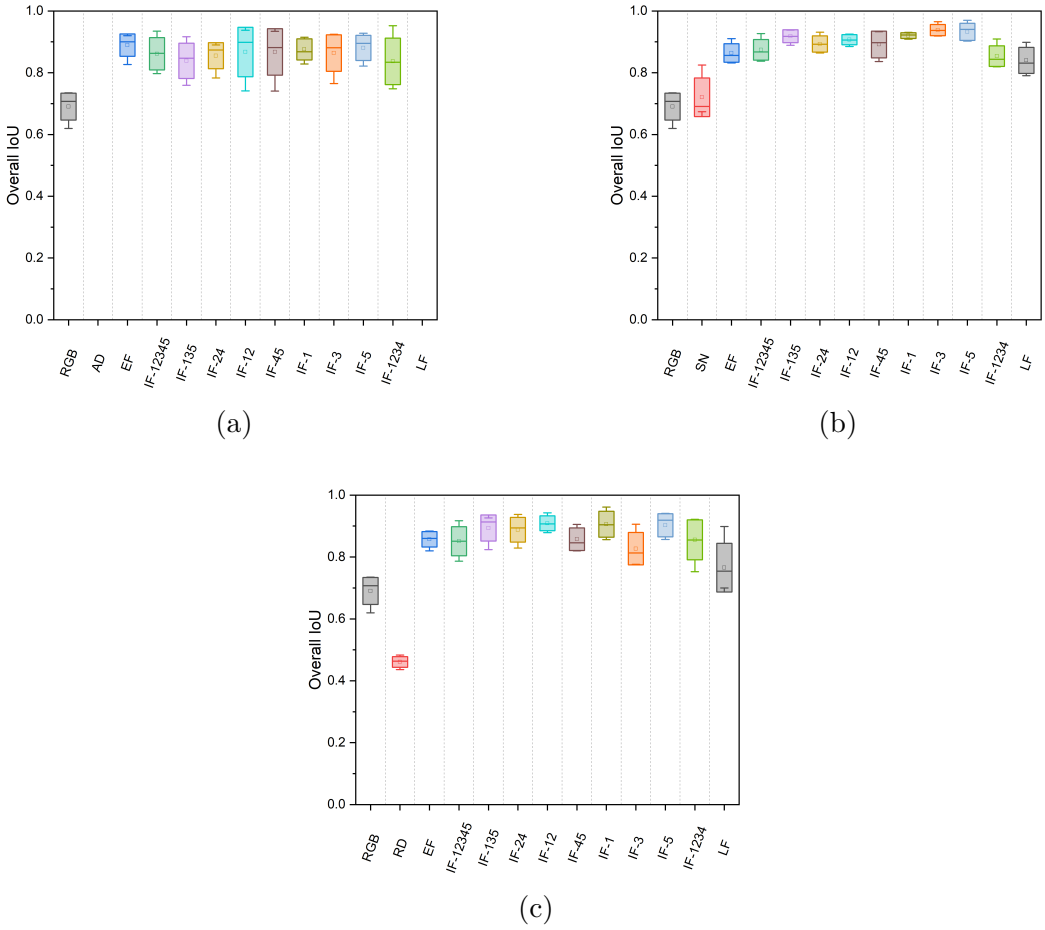
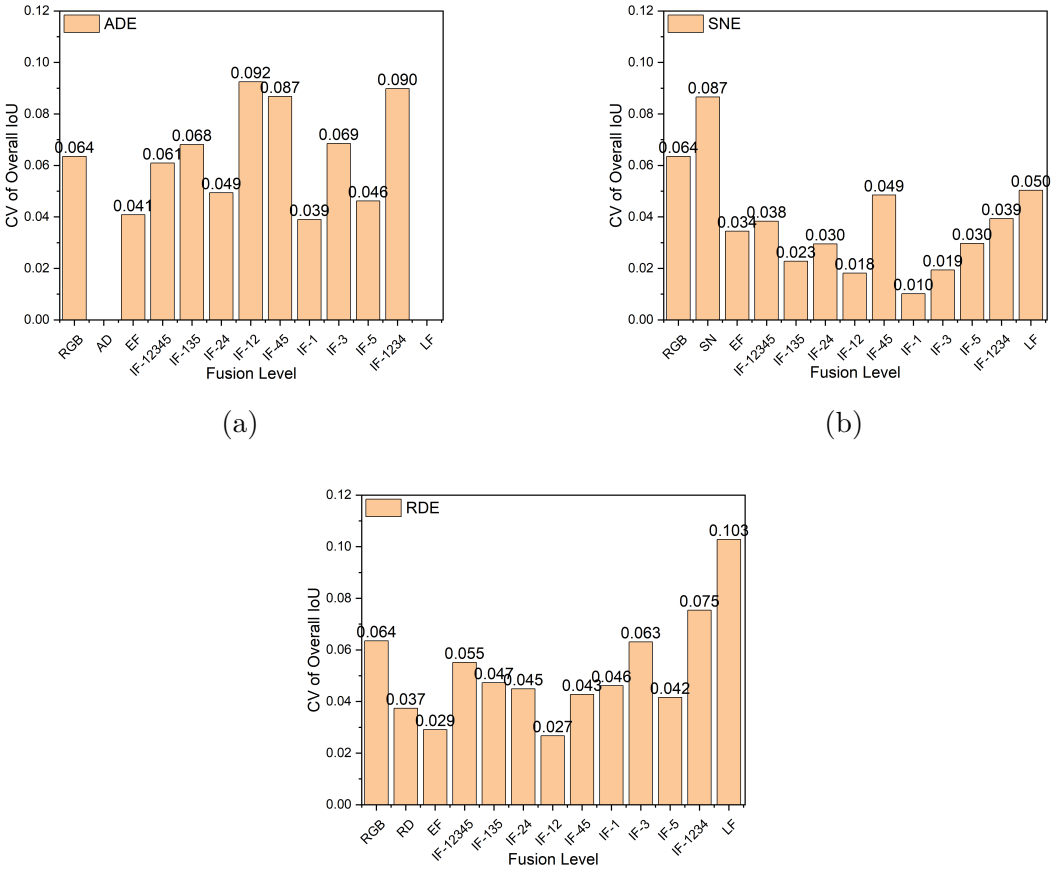


Figure 4.9. The overall IoU produced by cross-validation for - (a) Absolute depth-based encoding (ADE), (b) Surface normal-based encoding (SNE), (c) Relative depth-based encoding (RDE).



(a) (b)

(c)

Figure 4.10. Coefficient of variation (CV) of overall IoU for - (a) Absolute depth-based encoding (ADE), (b) Surface normal-based encoding (SNE), (c) Relative depth-based encoding (RDE).

substitute RGB data in presence of sensor noise. On the other hand, the network trained solely with SN data produced an accuracy higher than pure RGB-based model. This implies that SN data can not only provide complementary information, it can even be a substitute to RGB data. Therefore, in absence of proper lighting where RGB-based algorithms may fail, SN-based models can be leveraged for semantic segmentation of damage based entirely on geometric features without recourse to any kind of photometric information. Similar deductions can extended to RD data, however with significantly reduced efficiency.

The overall IoU of a robust fusion strategy should not only have a high mean, but it should have a small variance also. Therefore, the relative efficacy of various fusion strategies are evaluated in terms of the coefficient of variation (CV) given by the ratio of the standard deviation to the mean of a population. A lower value of CV indicates higher robustness. As observed in Figure 4.10, IF-1 exhibits the lowest CV for ADE, and therefore, it is designated as the most robust fusion strategy for this encoding technique. In the same way, IF-1 and IF-12 can be identified as the most robust fusion strategies for SNE and RDE, respectively. Therefore, it can be inferred that different evaluation criteria lead to different results vis-à-vis selection of the best fusion strategy. It can also be noticed that depth fusion does not always improve the robustness of RGB-based models. Overall, SNE performs better than RDE in terms of accuracy and robustness as observed in Figures 4.9 and 4.10. ADE, on the other hand, is the least accurate and robust of all the encoding techniques considered in this study.

However, the overall IoU does not provide a class-wise estimate of the performance of a learning algorithm. Therefore, the most robust fusion strategy identified for each depth encoding technique is exploited to evaluate the class-specific IoU values as shown in Figure 4.11. It is observed that a pure RGB-based model does reasonably well in segmenting a normal spalling. However, the performance drops significantly as the damage severity increases. This is further evidenced by the sample test cases presented in Figures 4.12 and 4.13. It is noticed that several pixels depicting concrete spalling with exposed rebars were wrongly labeled by the RGB-based network as buckled rebars, and vice versa. However, these mistakes were greatly minimized by using fusion-based approaches. It should be recalled that an RGB-based model relies only on color information. What depth fusion brings to the table

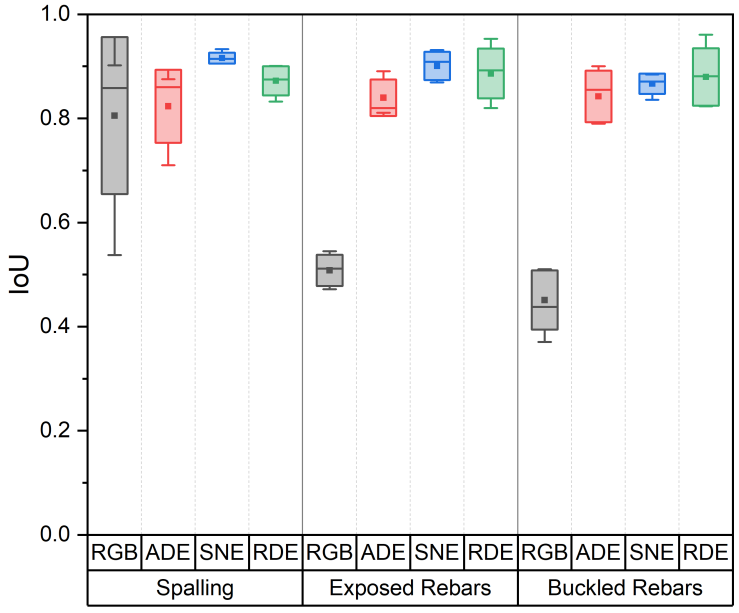


Figure 4.11. Class-wise IoU produced by fusion strategies exhibiting the least coefficient of variation (Figure 4.10) for different encoding techniques

is the crucial structural information that is necessary for distinguishing between straight and buckled rebars. Besides, it's worth noting that the damaged concrete surfaces in the backdrop of exposed and buckled rebars are texture-wise very similar. Therefore, any prediction based solely on the local context may err in correctly classifying the concrete (damaged) pixels in presence of exposed or buckled rebars. The depth fusion becomes helpful in such situations in better learning the global context as evidenced by Figures 4.12 and 4.13. This explains the greater dividend furnished by depth fusion in the case of more severe damage categories. Additionally, it should be noted that a pure SN-based model performs slightly better than an RGB-based model, which is also indicated by the overall IoU values reported in Figure 4.9. In this case, the discrepancies in the predicted label for pixels representing the damage boundaries can be attributed to the intricacies in fitting a local plane at the boundary points which is necessary for estimating the surface normals. In other words, a SN-based model does not predict the damage boundary very accurately. On the other hand, the semantic labels produced by networks incorporating a fusion of heterogeneous data are evidently far more accurate corroborating the favorable outcome of multi-sensor fusion.

Interpretability of DL algorithms is critical to scientific understanding and reliability of predictions. Therefore, recently there has been a heightened interest in the field of explainable AI, which seeks to justify model predictions in a human interpretable way. In the same spirit, this study aimed at obtaining a more granular view into the RGB-D fusion by probing the activation maps from the intermediate layers (Figure 4.14). A couple of activation maps from the first Conv+BN+ReLU block are shown in this figure for IF-1. It was observed that the RGB branch highlights certain features in the activation maps (bright regions). However, it also suppresses other features as denoted by the dark regions. On the other hand, the SN branch highlights those features that are largely suppressed by the RGB branch and suppresses those highlighted by the RGB branch. It is also noticed that the fused activations highlight more features than what is done by any single modality. This shows that RGB and depth data provide complementary information, which explicates the considerable performance boost afforded by the fusion of two modalities. It should be noted here that SN is only used as a test case in this illustration. However, the conclusions evenly hold true for other encoding techniques as well.

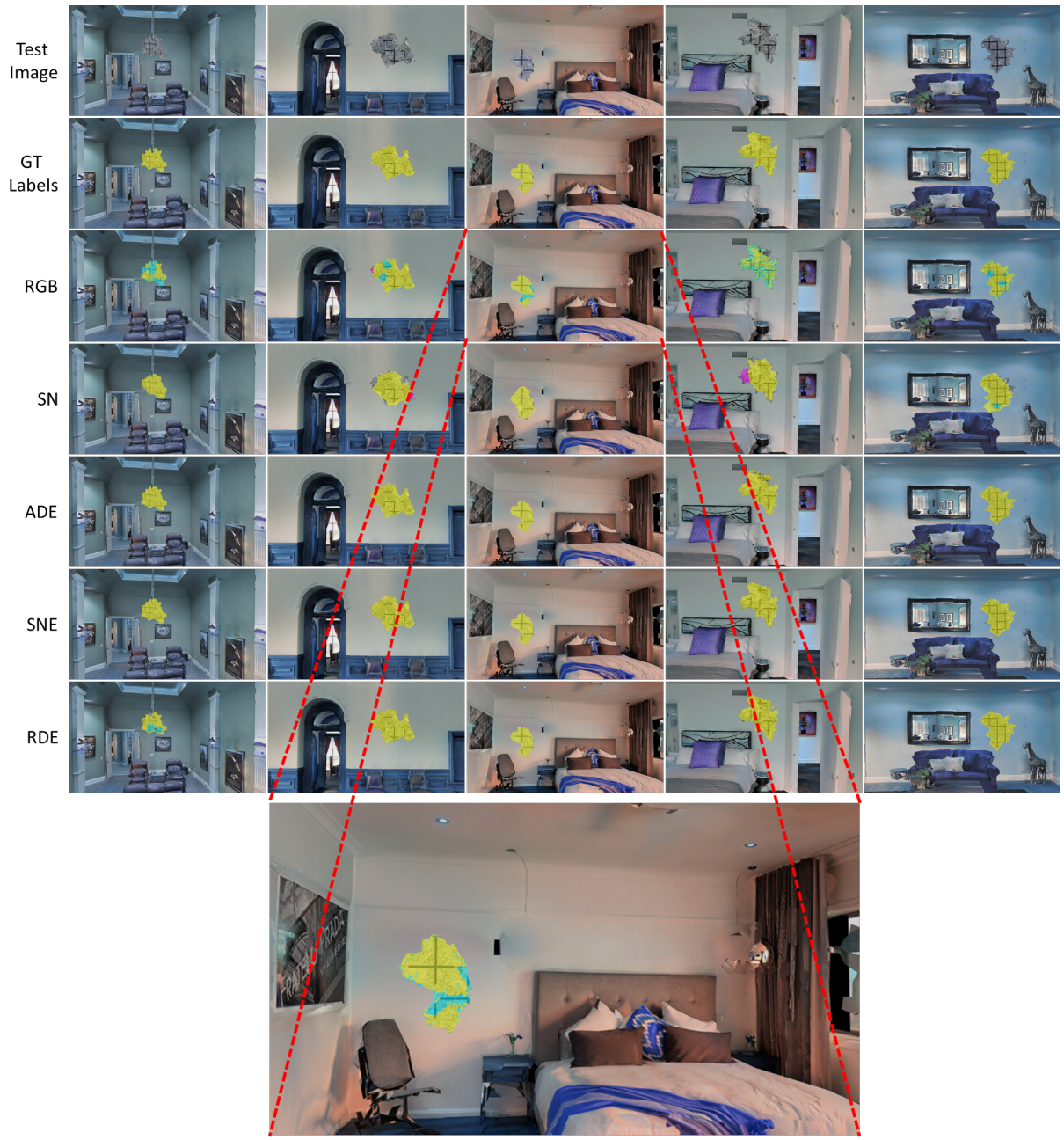


Figure 4.12. Qualitative segmentation results. The ground truth damage category is exposed rebars. Magenta color denotes spalling, yellow color denotes exposed rebars, cyan color denotes buckled rebars.

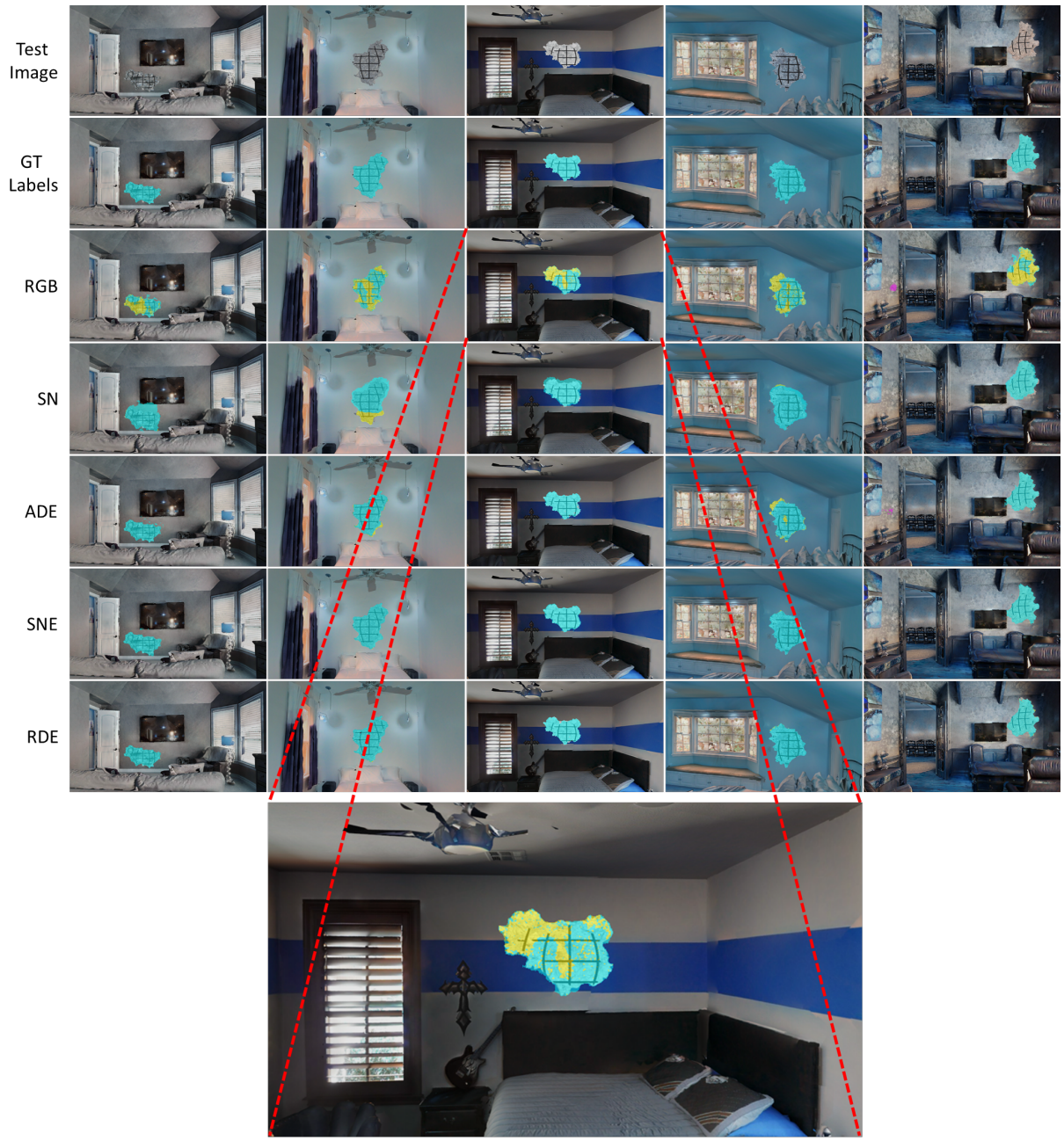


Figure 4.13. Qualitative segmentation results. The ground truth damage category is buckled rebars. Magenta color denotes spalling, yellow color denotes exposed rebars, cyan color denotes buckled rebars.

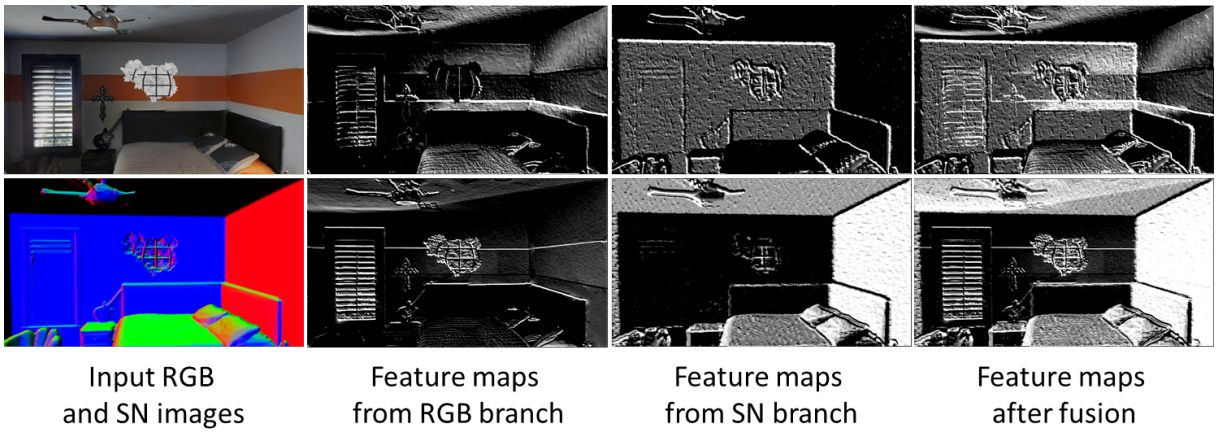


Figure 4.14. Visualization of features from the first Conv-BN-ReLU layers for IF-1. This figure shows that the fusion of RGB and depth activations highlights more features of the scene than what any single modality operating individually can do. In other words, RGB and depth data provide complementary information, the fusion of which leads to improved segmentation performance.

Validation of these important findings with real data was not practicable due to the lack of a publicly available damage database that contains depth information. Acquisition of real data was further inconvenienced by the scarcity of seismically damaged RC buildings in the local vicinity. As a makeshift arrangement, a first-generation HoloLens (Figure 4.15) was used to photograph a few damaged RC specimens, which were visually not quite comparable with the damage scenarios considered for training. The HoloLens [116] developed by Microsoft is a head-mounted augmented reality device which is equipped with RGB and depth sensors. Therefore, it is a perfect platform for collecting multi-modal inspection data. However, it should be noted that the RGB and depth images recorded by the HoloLens have different resolutions and fields of view. Therefore, an image alignment algorithm had to be invoked to synchronize the RGB and depth frames. The predictions of the trained models on these real data are presented in Figure 4.16, which reaffirm the superior performance of fusion-based models compared to the baseline RGB-based network. The overall poor performance of the models on real data can be attributed to the mismatch between the test specimens in real data and those in the synthetic data which were used for training of the networks. A more extensive study based on real RGB-D data is scope for future work.

Another increasingly important parameter in this age of edge computing and cloud computing is the processing time. The unmanned aerial systems used for inspection are often incapacitated by limited payload capacity to carry high-powered computing devices. Therefore, a real-time damage diagnosis requires that the chosen segmentation model is computationally less expensive. The computational cost is measured in this study in terms of the processing time needed to process a single image as shown in Figure 4.17. It should be noted here that the processing time depends greatly on the image resolution and specifications of the processing unit used for computation. An NVIDIA Quadro RTX 8000 GPU was used in this study, and the image resolution was 768 pixel $\times$ 432 pixel. It is observed that ADE is the least expensive of all encoding techniques regardless of the fusion strategy. The SNE requires a relative higher processing time, which can be attributed to the need of estimating SN map from the corresponding AD data as a part of data preprocessing. However, RDE requires the maximum processing time. This is due to the multi-stage approach of RDE. The first stage, which entails detecting a damaged region using the Faster RCNN algorithm,

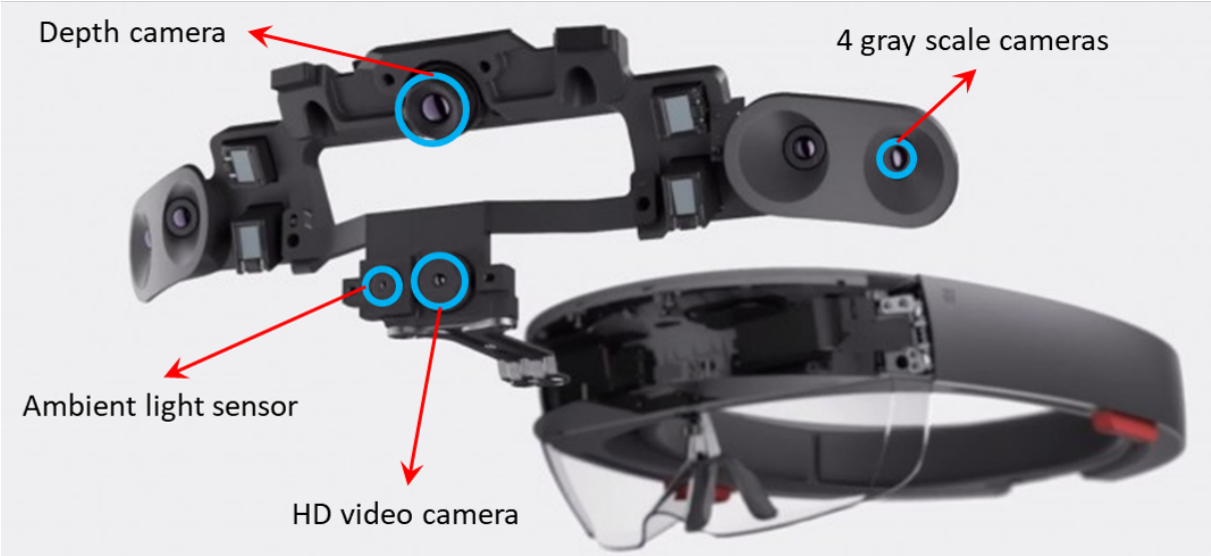


Figure 4.15. First generation Microsoft HoloLens. It is a head-mounted augmented reality device capable of RGB-D sensing.

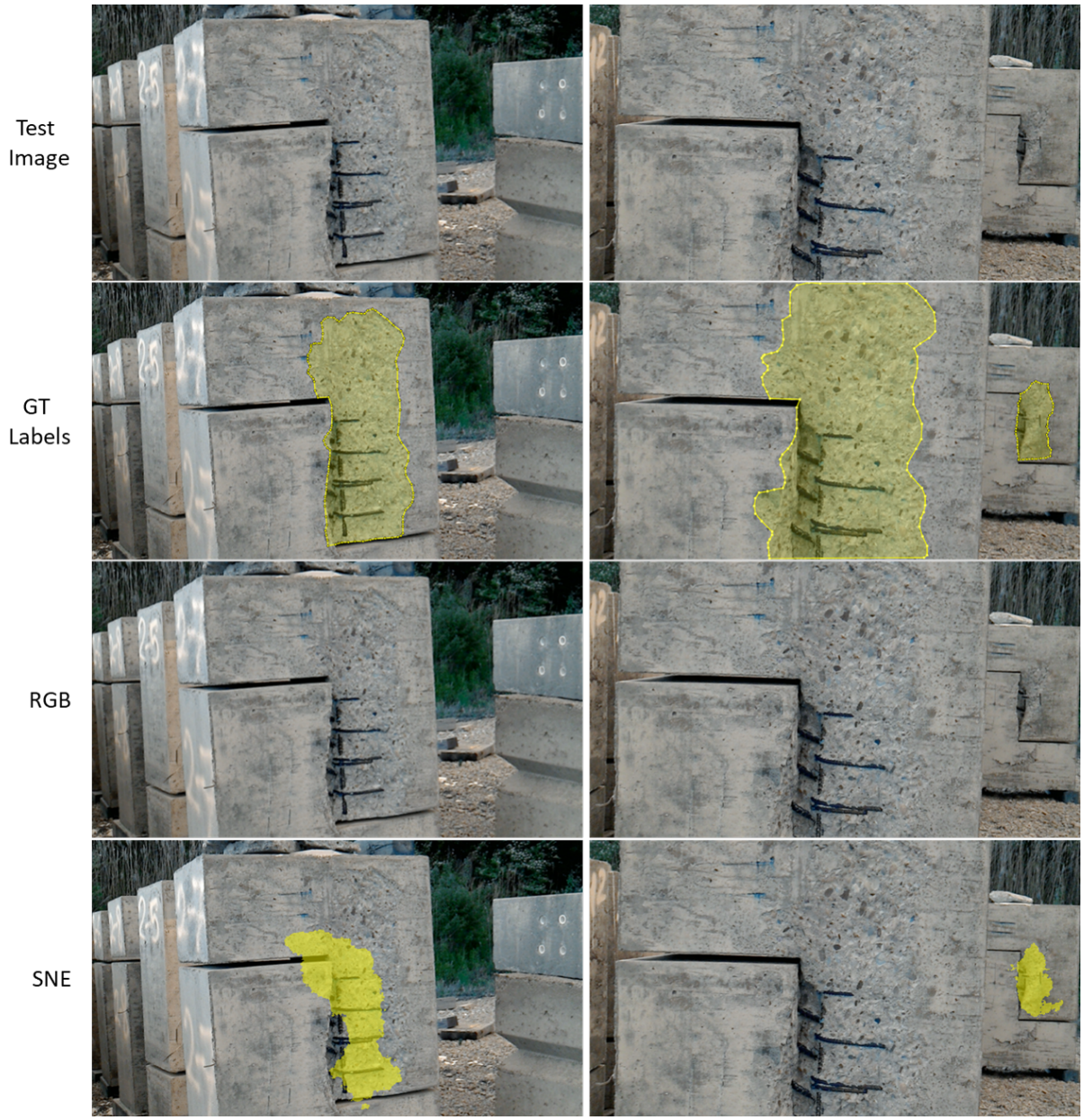


Figure 4.16. Qualitative segmentation results for real data collected by Microsoft HoloLens. The ground truth damage category is exposed rebars and is denoted by yellow color.

takes around 0.1265 seconds on an average to process a single image. The following stage, which seeks to fit a plane using the RANSAC algorithm and estimate the RD, consumes the maximum amount of time which is estimated to be 0.397 seconds per image on an average. It should be noted that this piece of computation relied solely on the CPU, and no GPU was used at this stage. The remaining inference time for RDE concerned semantic labelling of damages using the proposed FCN similar to ADE and SNE. Therefore, RDE is the least recommended for real-time condition assessment of civil infrastructures.

4.5.1 Damage Quantification

One of the advantages of semantic segmentation is that it lays the foundation for damage quantification. Depth sensors have been used in the past for quantification of concrete cracks [73] and spalling [94]. However, the volumetric quantification approach presented by Beckman et al. [94] did not incorporate perspective correction. Therefore, the technique proposed in this study will yield inaccurate results when the sensor plane is not parallel to the structural surface being investigated. The present study introduces a novel approach for volumetric quantification of concrete spalling which is robust against perspective distortion.

In this approach, the damaged area is first identified using the Faster RCNN algorithm. The equation of the plane representing the unbroken surface surrounding a damaged region is then determined (Figure 4.18a) using a RANSAC-based plane fitting algorithm as described in Section 4.3 of this chapter, and a RD map is obtained. Each pixel in the depth map is then back-projected on the object plane. To this end, a line is drawn passing through the optical center of the depth camera and the four corners of a pixel under consideration (Figure 4.18b). The equations of these lines can be easily obtained by utilizing the known coordinates of the pixel corners and the camera center, through which the lines pass. This is followed by the estimation of the intersection points of these four lines with the object plane forming a quadrilateral as shown in Figure 4.18b. This quadrilateral may not have the same shape as the pixel due to perspective distortion. The area of this quadrilateral multiplied by the RD at the pixel location yields the volume of damage at the given pixel. The summation of volumes calculated at all pixels will produce the total estimated volume

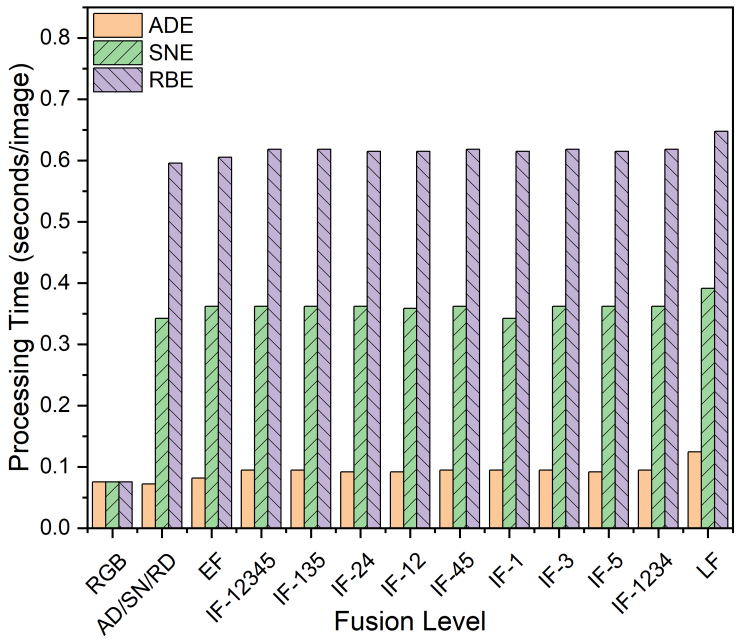


Figure 4.17. Processing time for different depth encoding techniques

of damage. An identical approach was followed to compute the actual volume, except that in this case ground truth information was used to identify the damaged regions without recourse to the Faster RCNN algorithm. The estimated damage volumes for different test cases compared favorably with the actual values as shown in Figure 4.19.

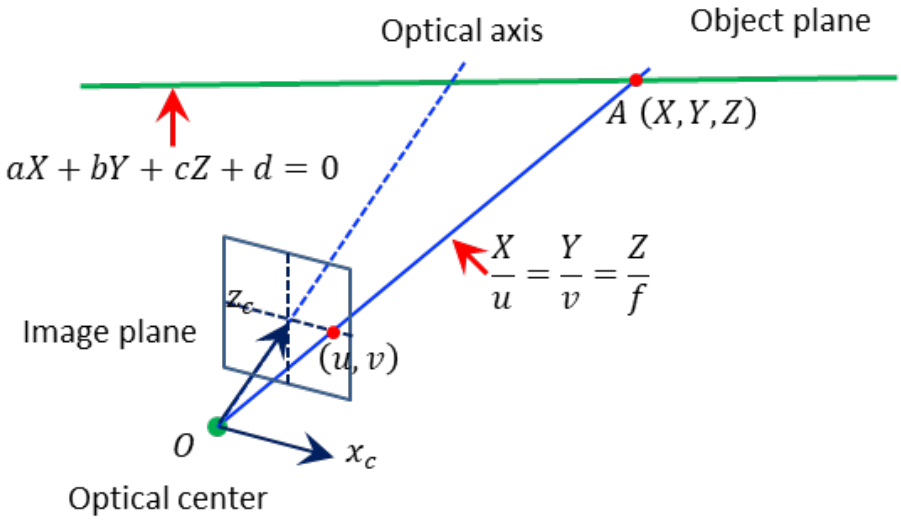
The existing condition assessment approaches do not involve measuring damage volume due to limitations in sensing capabilities. However, barrier-free access to low-cost consumer-grade depth sensors in recent years has increased the viability of volumetric quantification of damage, which will potentially lead to more accurate and robust decision-making in the future. In all probability, this will drive the emergency management agencies to update the existing inspection manuals to reap the benefits of such valuable technological advancements.

4.6 Addressing Practical Challenges of Depth Sensing

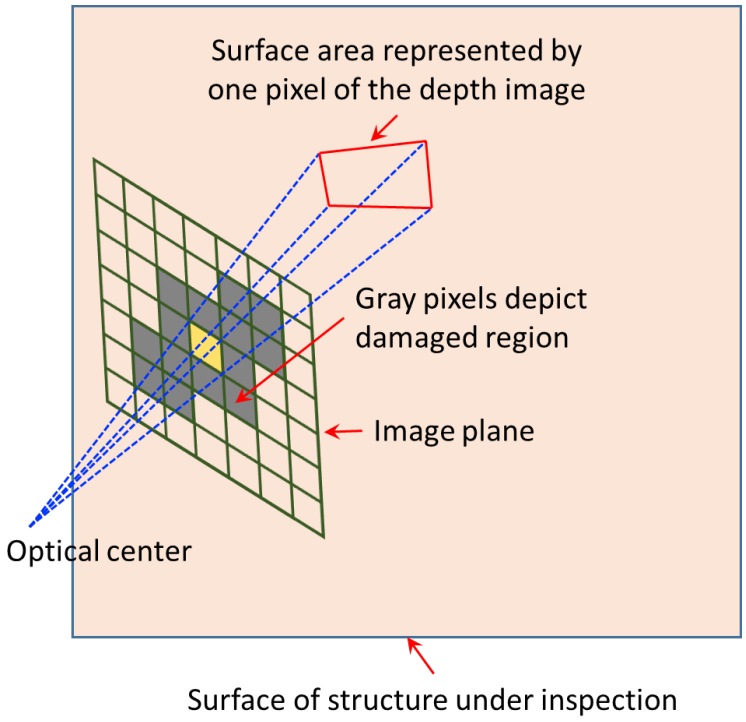
Depth fusion is not an unmixed blessing. Alongside the advantages, depth sensing also presents many practical challenges. Therefore, the feasibility of doing away with depth sensing at test time without compromising the segmentation performance is investigated in this study with the help of two recent developments in the field of vision-based multi-modal sensing, namely modality hallucination (MH) and monocular (single view) depth estimation (MDE).

4.6.1 Modality Hallucination

In MH, ED is used at training time as side information to produce a more informed test-time RGB only network. This technique is implemented in this study by developing on the same encoder-decoder network used earlier in this chapter. In this technique, an access to paired RGB and ED images is presumed at the training time. Apart from the usual RGB and ED branches, a third encoder, known as the hallucination branch, is also introduced (Figure 4.20a), which takes RGB images as input. A regression-based hallucination loss is introduced between paired hallucination and ED mid-level activations to facilitate an efficient information sharing between the two modalities. This loss is minimized alongside a standard supervised loss over the class labels, ensuring that the mid-level convolutional



(a)



(b)

Figure 4.18. Volumetric damage quantification

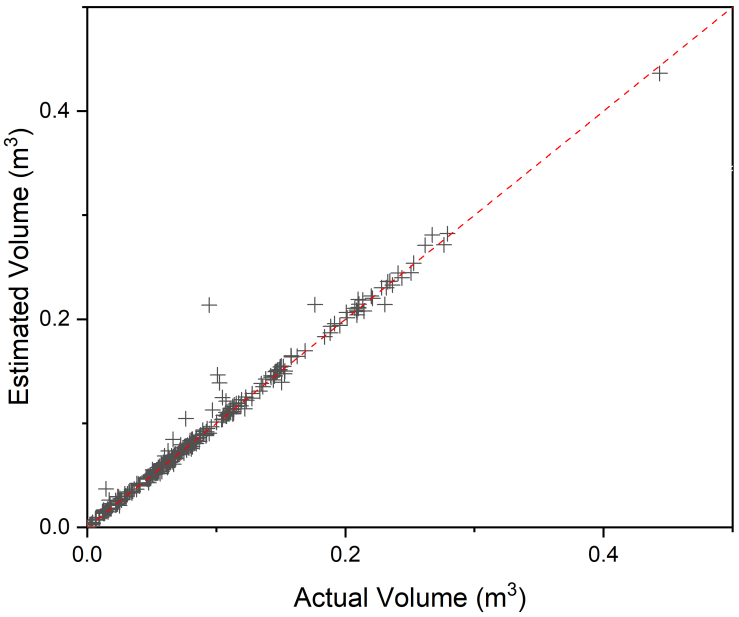


Figure 4.19. Comparison of estimated and actual damage volumes

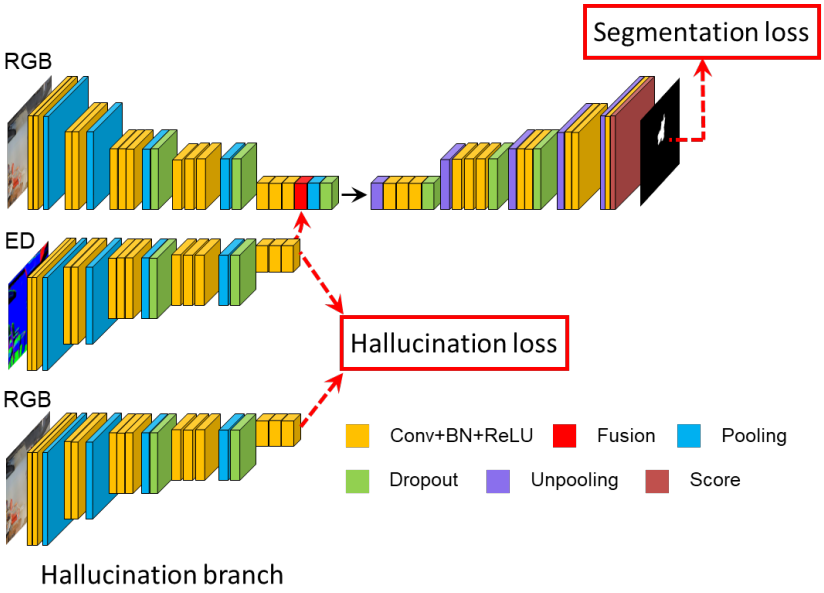
features learned by the hallucination and ED branches mirror each other. In consequence, the ED branch becomes redundant at the end of the training process. Because the same mid-level features, which were hitherto generated by the ED branch, can now be hallucinated by the hallucination branch using RGB data as network input. Thus, at test time, the ED branch can be discarded, and the mid-level activations from the hallucination branch can be fused to the RGB branch to emulate a multi-modal fusion (Figure 4.20b). This gives rise to a more informed test-time RGB-based network which significantly outperforms a standard benchmark model trained solely on RGB data. This eliminates the need for depth sensing at test time without any appreciable loss of segmentation accuracy.

4.6.2 Monocular Depth Estimation

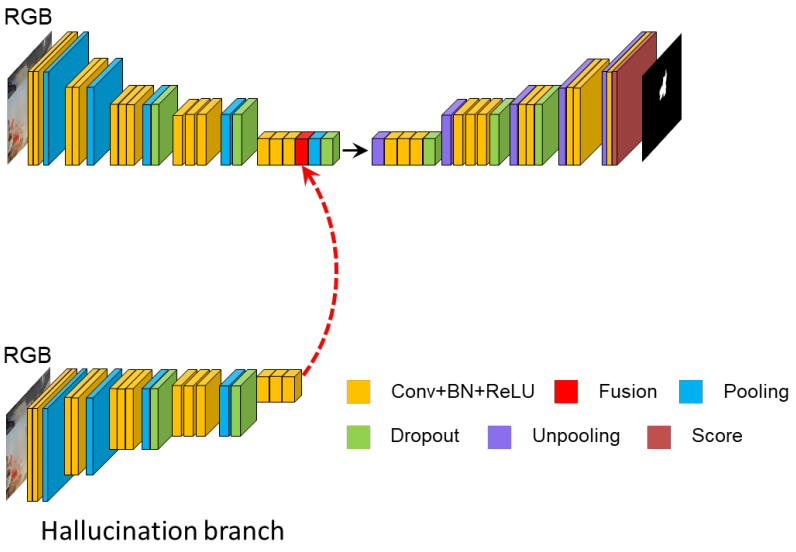
The goal of MDE is to predict pixel-wise depth values corresponding to a given RGB image. Traditional depth estimation methods such as structure from motion [117]–[120] and stereo matching [121]–[123] rely on multiple views of a scene to generate a sparse depth map. However, many real-time inspection applications require depth map to be estimated from a single viewpoint. The recent developments in DL-based computer vision techniques have shown great promise of enabling this challenging task by predicting a dense depth map from a single frame RGB image in an end-to-end manner. This study explored two different approaches to this end based on convolutional neural network (CNN) and generative adversarial network (GAN). The reconstructed depth maps can then be paired with the corresponding RGB images to be used as inputs for the fusion-based segmentation models presented earlier in this study.

CNN-based approach

A standard encoder-decoder-based CNN with skip connections (Figure 4.21) is used to predict detailed high-resolution depth maps from single frame RGB images. The encoder is borrowed from a DenseNet-169 architecture [124] pre-trained on ImageNet dataset [113]. The decoder, on the other hand, comprises a series of up-sampling layers. The baseline architecture is adopted from Alhashim and Wonka [125] with a few modifications. In the



(a) At training time



(b) At test time

Figure 4.20. Modality hallucination architecture. It is trained to counterfeit intermediate depth features from input RGB image, which makes depth sensing redundant at test time.

original study, the resolution of the final output depth maps was half the input resolution. However, the fusion strategies proposed in this study require that the input RGB and depth images should have the same resolution. To address this specific need, this study appended an additional upsampling layer at the end of the network to ensure that the output resolution matches that of the input. The predicted depth values are regressed to ground truth depths by minimizing a composite loss function consisting of an L1 loss defined on the depth values, an L1 loss defined over the gradients of depth image, and a structural similarity loss. The efficiency of this approach is discussed in Section 4.6.3.

GAN-based approach

A number of studies [126]–[130], on the other hand, resorted to GAN for MDE. A GAN consists of a pair of neural networks known as the generator and the discriminator, which fight with each other. The generator is like a counterfeiter who tries to generate some fake depth images, and the discriminator is like the cop who tries to catch the counterfeiter. In the training phase, the generator becomes better and better at producing more realistic depth images until it can produce a perfect depth image, which fools the discriminator into believing that it is a real image. The same encoder-decoder network described in Section 4.6.2 is used in this study as a generator to produce some artificial depth maps, which were then classified by a discriminator as real or fake (Figure 4.22). The discriminator in this study, which facilitated this adversarial training, was adopted from the classical CycleGAN paper [131]. The performance of the GAN-based approach is described in the following section.

4.6.3 Results

This section presents the results of CNN- and GAN-based MDE. Traditional DL algorithms require that the input data are suitably normalized before being fed into a DL model. Therefore, the ground truth and the estimated depth values were normalized between 0 and 1 in this study before computing the depth estimation accuracy. After this normalization, the estimated depth values were compared with the ground truth depths, and the average root

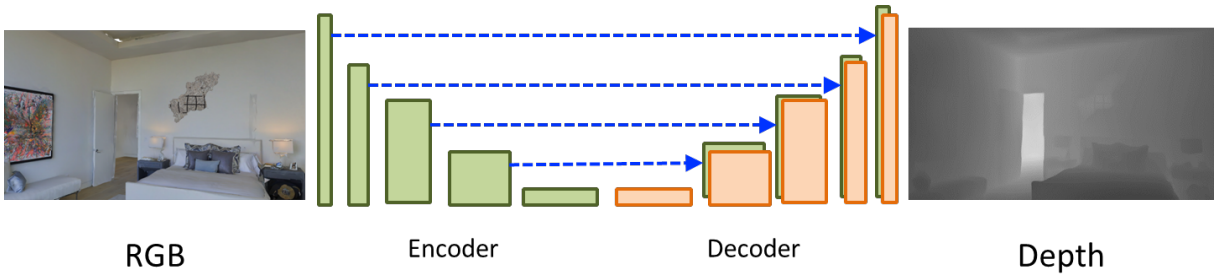


Figure 4.21. CNN-based monocular depth estimation

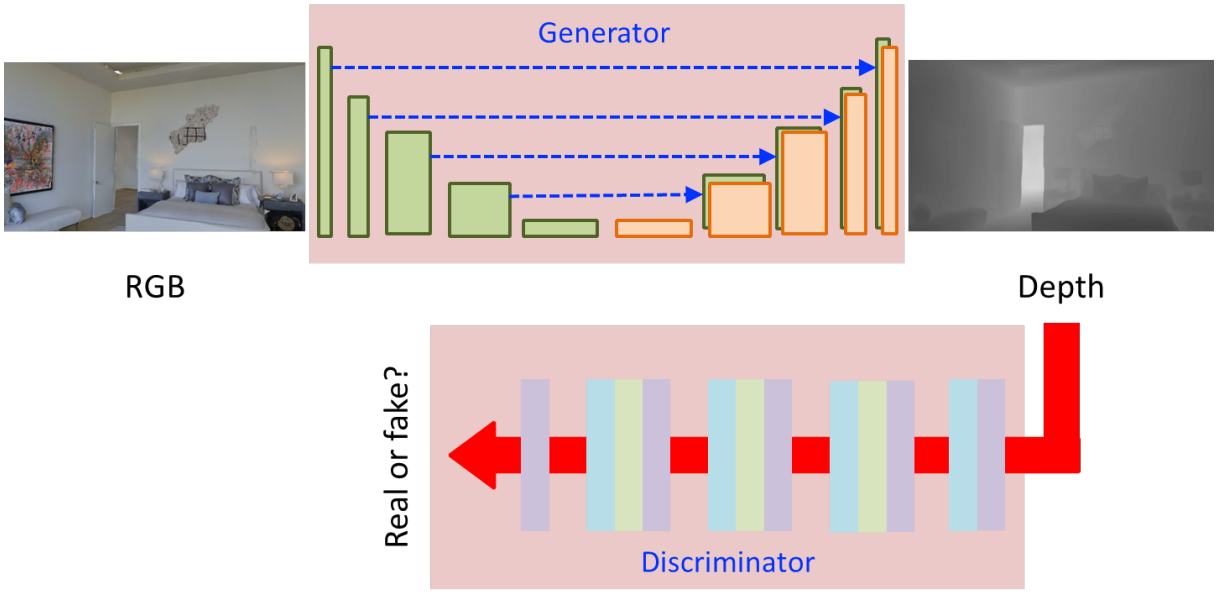


Figure 4.22. GAN-based monocular depth estimation

mean square errors for five-fold cross-validations were observed to be 0.0435 and 0.0452 for the CNN- and GAN-based approaches, respectively. This indicates that adversarial training was not of any significant help, and therefore was not considered for any subsequent analysis. A few examples of the depth maps generated by the CNN-based approach are shown in Figure 4.23 side by side with the corresponding RGB and ground truth depth images to demonstrate the efficiency of this technique.

The main purpose of invoking MH and MDE was to create proxies for real depth sensing at test time. Therefore, the efficiencies of these techniques were evaluated in terms of accuracy and processing speed, as shown in Figures 4.24 and 4.25, respectively. It is observed in Figure 4.24 that, in the case of ADE, MH and MDE have comparable accuracies, both being in the same ballpark with the measured depth (MD)-based approach. In the case of SNE, a 4% drop in the accuracy was noticed vis-à-vis MD when MH was used. However, this accuracy is still streets ahead of that of a single-modality RGB-based model. Also, MH demonstrated a slight edge over MDE in terms of segmentation accuracy for this encoding technique. On the other hand, in terms of processing speed, it was observed (Figure 4.25) that MH offers a major advantage for both ADE and SNE. It requires a processing time that is even lower than an MD-based model and is at par with a pure RGB-based model. It is particularly advantageous for SNE, where considerable time is expended in SN estimation from raw depth measurements. This step becomes inessential when MH is used. It leads to a win-win situation on all counts as it increases the accuracy at no additional cost of processing time. However, the MDE-based technique requires considerably higher processing time, more so in the case of SNE. This can be attributed to the two-stage process involved in this approach, namely depth estimation and semantic segmentation. Therefore, in the overall analysis, it can be concluded that MH has a comparative advantage over MDE in terms of both accuracy and processing speed. It can ably compensate for the lack of depth data at test time. This implies that depth-sensing at test time is not indispensable. On the contrary, depth sensing can be surrogated at test time by employing state-of-the-art MH techniques. It is believed that this is a significant addition to the existing knowledge base and will go a long way to enhance the efficiency of robotic inspection in time to come.

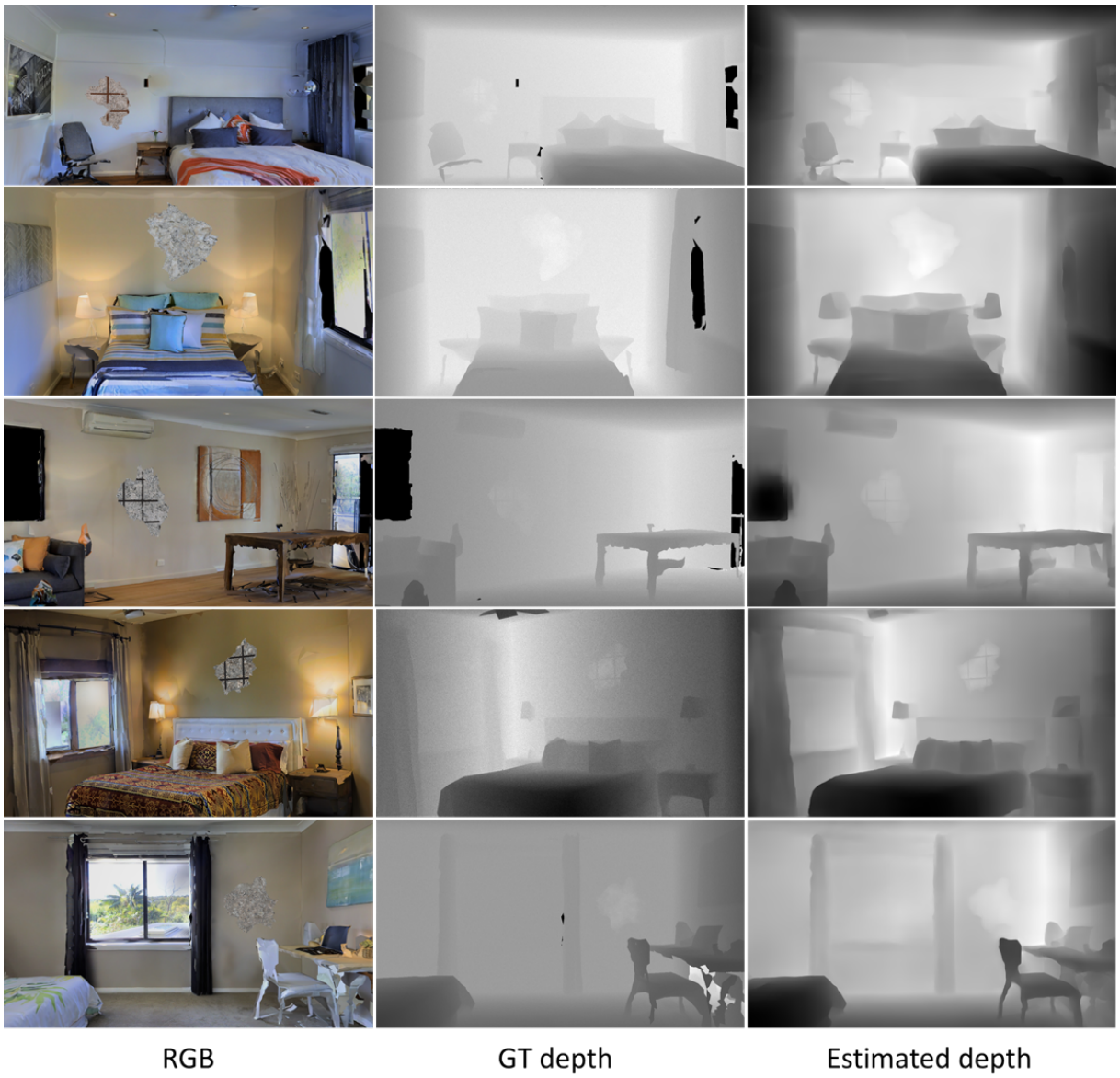


Figure 4.23. Examples of monocular depth estimation using CNN-based approach

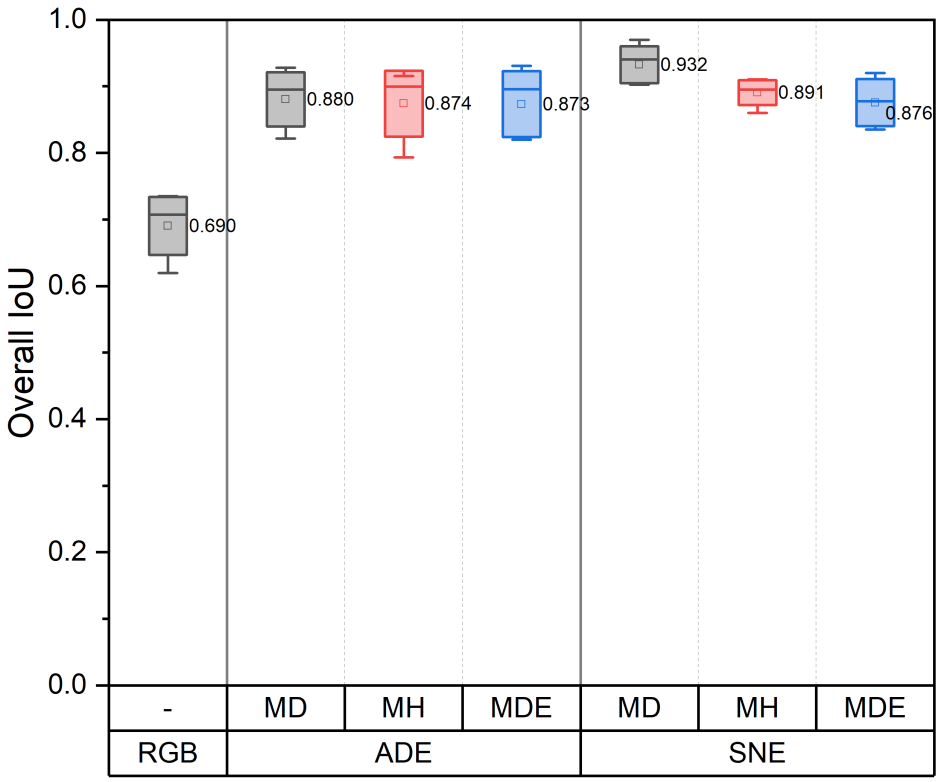


Figure 4.24. Accuracy modality hallucination (MH) and monocular depth estimation (MDE) as compared to measured depth (MD)

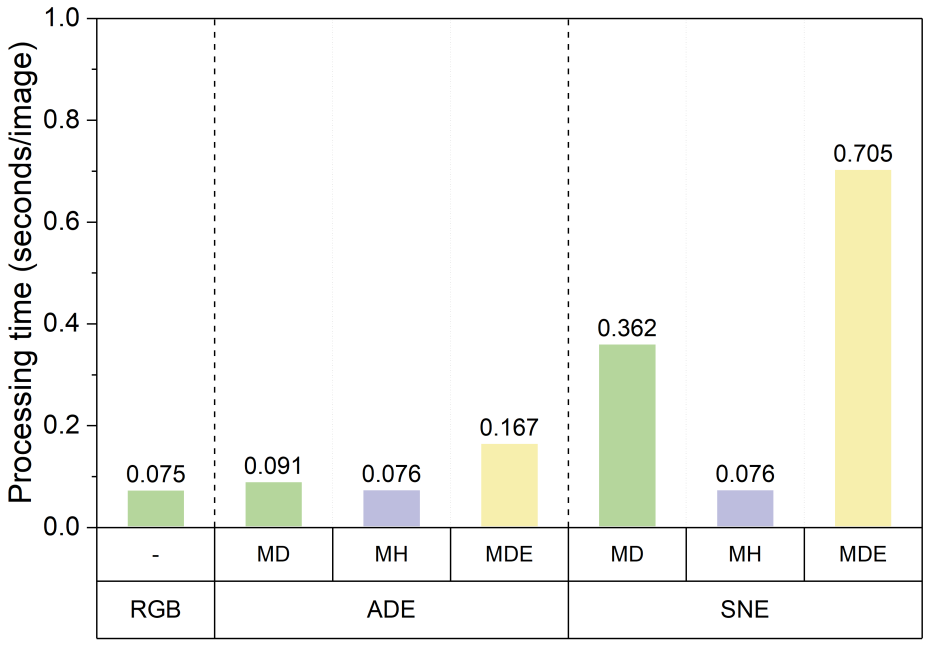


Figure 4.25. Processing time for modality hallucination (MH) and monocular depth estimation (MDE) as compared to measured depth (MD)

4.7 Conclusions

This study is perhaps the first of its kind which investigates the effect of depth fusion on the performance of deep learning-based multi-class damage segmentation framework. A synthetic database is generated using state-of-the-art computer graphics techniques containing three different damage categories which are commonly observed in RC structures subjected to extreme loading. Several experiments are conducted which suggest that depth-fusion considerably enhances the performance of RGB-based segmentation models. Various encoding techniques are considered to represent depth data, including RDE which is a novelty of this work. The SNE was observed to outperform the ADE and RDE in terms of accuracy and robustness. Additionally, various strategies are explored for fusing depth information, and the best strategy is identified for each encoding technique. The benefits from depth fusion is found to be higher for exposed and buckled rebars than normal spalling. The ADE requires the least processing time, followed by SNE and RDE. On the whole, it can be inferred that the structural information provided by depth data complements the color information embedded in RGB images leading to an improved segmentation accuracy. This study also explored two surrogate approaches based on MH and MDE to get rid of depth sensing at test time without letting go of its invaluable benefits and found MH to be more efficient. A novel volumetric damage quantification approach is also proposed which is robust against perspective distortion. It is believed that an effective implementation of the depth fusion techniques presented in this study will prove to be a major breakthrough in the realm of vision-based autonomous condition assessment of civil infrastructures. Validation of the proposed approach with real depth data is a scope for future work.

5. SUMMARY AND CONCLUSIONS

This dissertation focused on expanding the frontier in vision-based autonomous structural inspection and condition assessment. A number of studies are undertaken with an aim to develop fully autonomous robotic inspection and monitoring systems. First of all, a deep learning-based approach is proposed for quick post-disaster reconnaissance of reinforced concrete buildings. Faster RCNN algorithm is employed to detect multiple damage categories, namely, surface crack, spalling, severe damage with exposed rebars and severely buckled rebars. State-of-the-art CNN models such as Inception v2, ResNet-50, ResNet-101 and Inception-ResNet-v2 are exploited as backbone architectures. The networks were trained and evaluated on image data collected from several past earthquakes such as Nepal (2015), Taiwan (2016), Ecuador (2016), Erzincan (1992), Duzce (1999), Bingol (2003), Peru (2007), Wenchuan (2008), and Haiti (2010) earthquakes. It is observed that Inception-ResNet-v2 outperforms other networks in terms of mean average precision. Moreover, it also noted that accuracy is inversely proportional to the processing speed of the detection algorithms.

Additionally, a novel computer vision-based approach is proposed to benefit from the valuable information hidden in historical inspection data. Strategies are proposed for autonomous exploration into erstwhile inspection data in search of correspondences, view synthesis from multiple correspondences and alignment to the current scene under consideration, localizing damage in the reconstructed scenes from the past, segmenting damage, and finally quantifying the damage to extract necessary information and derive meaningful conclusions, after a damage is detected in the current data set. Temporal evolution of the damage is graphically presented facilitating easier interpretation in addition to predictive and quantitative evaluations. Cracks on concrete surface are used as a case study for validation of the proposed approach. However, it can be extended to other damage categories such as spalling and corrosion which is a scope for future work.

Furthermore, this study aimed at unlocking the next level in vision-based inspection of civil infrastructure by incorporating depth fusion into a CNN-based damage diagnosis framework. Three different damage categories which are commonly observed in RC buildings are considered in this study, namely, spalling, spalling with exposed rebars, and severely

buckled rebars. Computer graphics software is exploited to synthetically generate a database of synchronized photometric and depth images to be used for the training and validation of an encoder and decoder-based semantic segmentation model. Three different encoding techniques are explored to represent the depth data. Additionally, various schemes for fusion of RGB and depth data are investigated to identify the best fusion strategy. The results of this study indicate that the performance of vision-based damage segmentation algorithms can be significantly improved by fusion of multi-modal data, leading to more accurate and robust inspection and enhanced infrastructure resilience. On top of that, this study proposed surrogate strategies to dispense with depth sensing at test time without compromising on the segmentation accuracy.

5.1 Future Work

This dissertation tackled only a few of the knowledge gaps existing in this important area of research. There are many other critical research questions which are yet to be addressed. The semantic segmentation of earthquake induced damages in RC buildings is one such area which should be taken up by future studies. In segmentation-based algorithms, each pixel in an image is classified and labeled according to the class it represents. Therefore, it has the ability to predict the shape of a damaged area more accurately than bounding box based approaches such as Faster RCNN. It may immensely benefit vision-based structural inspection, given that the shape of an damaged region is a powerful discriminator among different damage categories relevant to earthquake reconnaissance of RC buildings. For instance, shear cracks are preeminently diagonal, while flexural cracks usually spread in vertical or horizontal direction. This will facilitate finer level of categorization of various damages commonly observed in RC buildings post earthquake events. Additionally, it will help quantifying the severity of damage through autonomous evaluation of crack thickness and spalling area. Future studies should also explore the possibility of improving the detector performance by implementing Bayesian data fusion as proposed by [32]. Aggregating the detection scores of a damaged region photographed from disparate camera positions may eliminate some of the false detections leading to an improved detection accuracy. Future

studies should also focus on the practical implementation of this detection algorithm by integrating it with UAVs or inspection robots for real hands-on experiments.

Estimating the service life of a structure is important for the sake of scheduling future maintenance. Mechanics-based models are widely used and usually most reliable in this matter [132]–[134]. However, in absence of proper analytical model, statistical data driven approaches are adopted which rely on observed data from the past [135]–[137]. Probabilistic evaluation of historical damage evolution data helps predict the expected timeline for a specified serviceability limit state. This calls for establishing the chronology of a damage by exploring an archive of visual inspection data, which was not thoroughly studied by researchers in the past. The present study will potentially fill that knowledge gap and will make it possible to anticipate the remaining life of a civil infrastructure system by exploiting a statistics-based prognostic model, the detailed investigation of which is beyond the scope of the present study.

Besides, estimation of loss due to possible seismic events is an important interest area for planners, government organizations and insurance agencies. It helps them in disaster planning, formulating risk reduction policies, decision making on retrofit and mitigation strategies, and in calculating insurance rating. Evaluating the probability of reaching or exceeding a damage state given a specific value of intensity measure is a prerequisite for seismic loss estimation and risk assessment of infrastructure systems. This probability represented graphically is known as the fragility curve. Professional judgment provided by a panel of experts is one of the commonly used approaches for generating fragility curves, even though it lacks credibility on account of being subjective and dependent on expertise of individual experts. The damage prognostication approach alluded in this chapter can open up a new avenue of research in the direction of image-based fragility curve generation exploiting the chronological information embedded in archival data.

This dissertation will also provide an impetus to multi-modal inspection. Validation of the proposed fusion-based segmentation approach with real depth data is scope for future work. Recent developments in the field of adversarial domain adaptation [138] can be exploited to ensure that the proposed network which is trained on synthetic data, performs reasonably well on real data during actual inspection. Multi-sensor information fusion

leveraging state-of-the-art infrared and hyper-spectral cameras is another promising area of inquiry. The scope for future research also extends to updation of the existing inspection manuals to include volumetric quantification of damage severity capitalizing on the depth perception.

A. TIME-BASED AUTONOMOUS MONITORING OF CRACKS ON THE METALLIC WHEELS OF NASA's MARS EXPLORATION ROVER

A practical application and extension of the vision-based approach presented in Chapter 3 for estimation of damage chronology is described in this section. A Mars exploration rover called Curiosity (Figure A.1) was launched in December 2011 as a part of the National Aeronautics and Space Administration's (NASA) Mars exploration program which landed on the Mars surface in August 2012. The primary goal of the rover was to explore the climate, geology, and presence of life-supporting environments on Mars. The mission was initially planned for two years and subsequently it was extended indefinitely. The rover, which has been operational since then, has recently started developing some cracks on its wheels (Figure A.2), and to make matters worse, the cracks are expanding over time. It can be noted in this context that the failure of a wheel will imply impairment of the whole system, as the rover will no longer be able to move on the Mars surface. Therefore, the scientists at the Jet Propulsion Laboratory (JPL) of NASA are keen to track the temporal evolution of these cracks on the rover wheels. There are a number of cameras which are mounted on the rover for environmental sensing. The cameras, once in a while, become active to capture images from different viewpoints. These images are timestamped and saved in specific folders. The standard operating procedure required the human operators at the JPL to manually browse through all the data directories to identify the relevant images facilitating a visual interpretation of the evolving nature of a crack under investigation. Needless to say that such a manual process is labor-intensive and time-consuming. This study, therefore, proposed a computer vision-based approach to automate this process.

The approach proposed herein is akin to one presented in Chapter 3. However, the current problem was more challenging considering that the wheel surface is not flat, making homography-based techniques less appropriate. In addition to that, the background and the lighting conditions are changing continuously as the rover moves on the Mars surface, adding to the complexity of the problem. In view of these challenges, three major changes are incorporated into the original algorithm presented in Chapter 3. In the original approach,

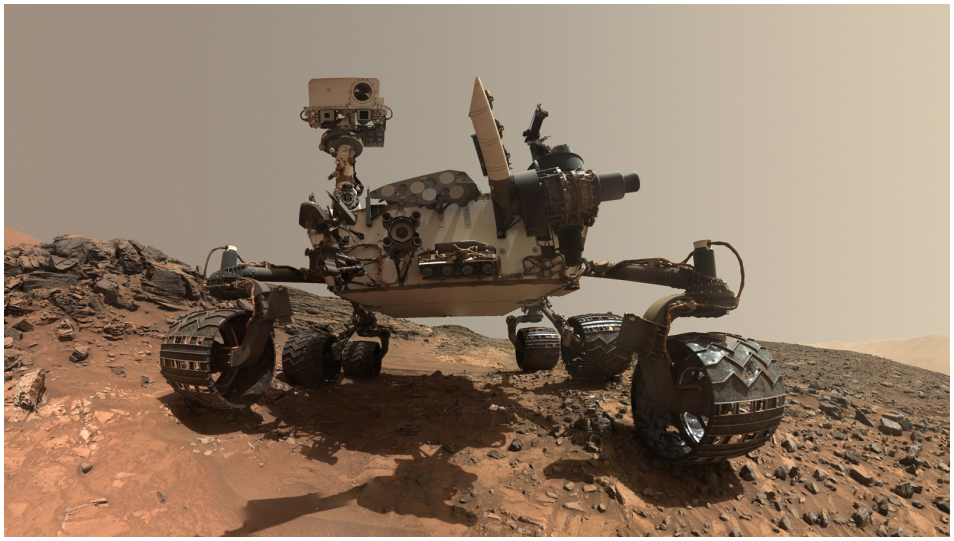


Figure A.1. Curiosity: NASA's Mars exploration rover

Figure A.2. Cracks on the rover wheel

the reference image used for feature matching and correspondence identification was the best correspondence from the immediately preceding dataset (arranged in the reverse order). However, in the revised approach, multiple reference images were considered comprising the best correspondences from the previous two data acquisition rounds (arranged in the reverse order) and the maiden initiatory reference image (Figure A.3). The sum total of matched features vis-à-vis all the three reference images was used as an evaluation metric to identify the best correspondence from a large pool of candidate images. Moreover, to ensure that no key point from outside the wheel surface is considered for feature matching, a region of interest is manually demarcated at the outset on the initial reference image (Figure A.4). The corresponding regions on relevant images from previous data acquisition rounds are automatically determined by homography-based warping of the original interest region. The feature matching exercise considered only those key points located inside the delimited interest regions, and any key points lying outside the regions were ignored. It was further observed that the identified best correspondence was not ideal in many cases, particularly when it contained the concerned crack region very close to a boundary of the image. Therefore, two candidate best matches were identified for each data set, and the one that contained the crack region closer to the image center was designated as the best correspondence (Figure A.5). Apart from these three modifications, the remaining procedure is identical to the approach elaborately discussed in Chapter 3, and the same is not repeated here.

Despite several challenges, it was observed that the autonomous approach developed in this study could accurately identify the appropriate corresponding images for a crack region of interest. A collage of best correspondences identified from historical visual data is shown in Figure A.6 for a test case under consideration. A lucid depiction of the temporal changes in a crack, as exhibited in this figure, will immensely benefit the scientists at JPL to efficiently monitor the deteriorating condition of the rover wheels.

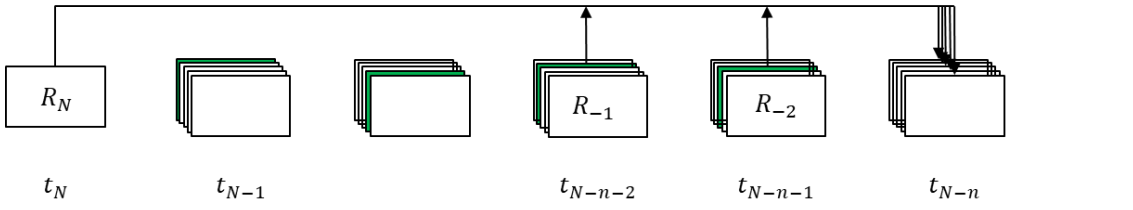


Figure A.3. To identify the best correspondence from a given data acquisition round, three reference images are considered, which include the initial reference image and the best correspondences from the immediately preceding two data sets (arranged in the reverse order). The evaluation metric used for identifying best correspondence is the total number of matched key points vis-à-vis all three reference images.

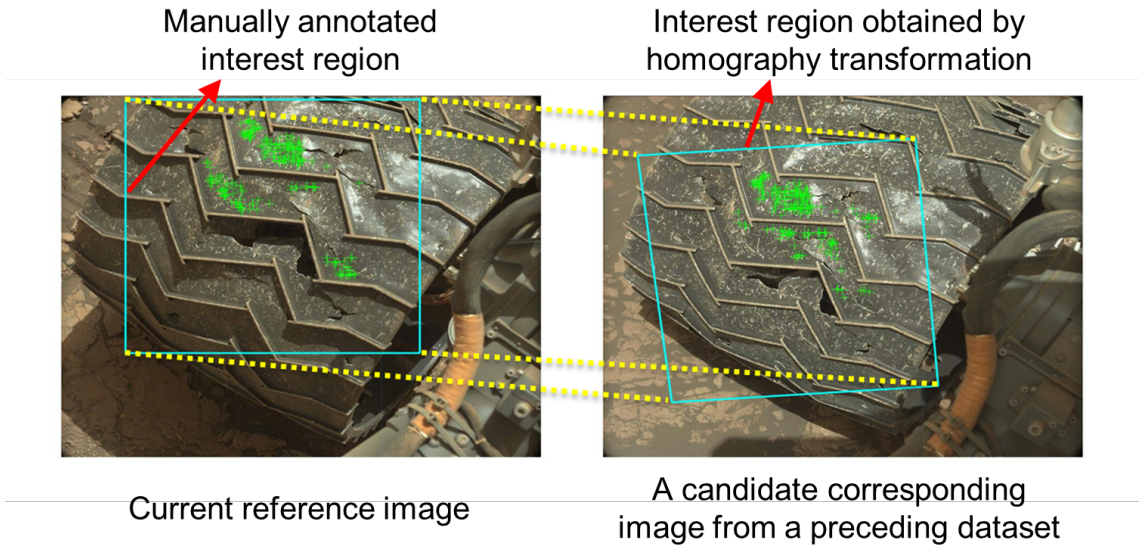


Figure A.4. To exclude all matched key points from outside the wheel surface, a region of interest was manually delimited in the initial reference image. The corresponding regions in the subsequent candidate correspondences are estimated by homography transformation of the concerned interest region.

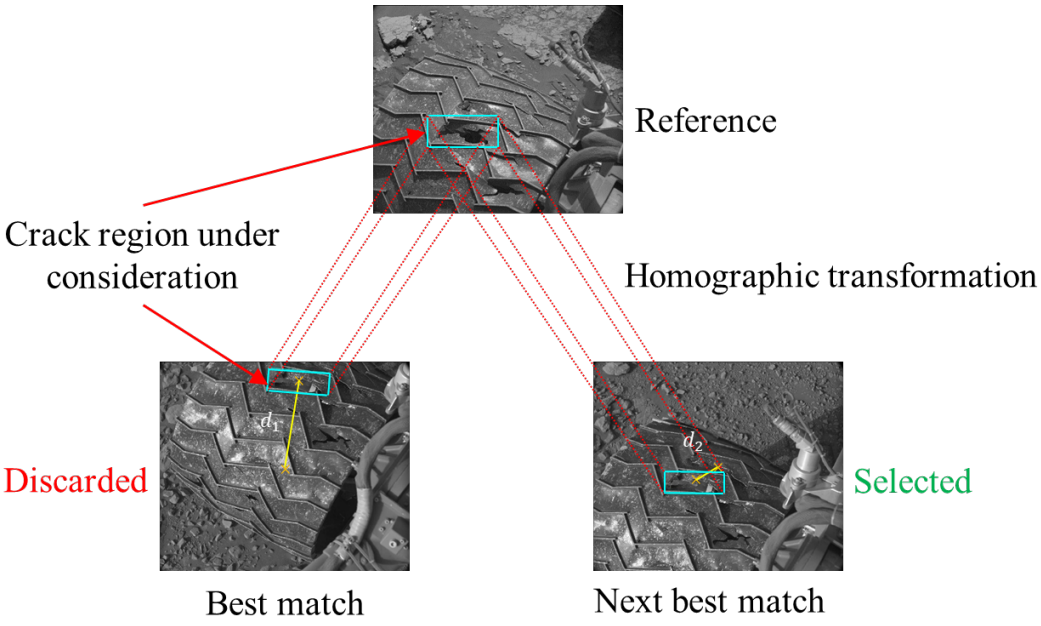


Figure A.5. The crack region in the identified best correspondence is often in the vicinity of the image boundary. This issue is addressed in this study by identifying two candidate best correspondences and then selecting the one in which the crack region is closer to the image center.

Figure A.6. A collage of identified corresponding images from various data acquisition rounds. The figure clearly depicts the time evolution of a crack under investigation in the reverse order. The figure legends represent the line-up of data acquisition rounds.

REFERENCES

[1] R. S. Adhikari, A. Bagchi, and O. Moselhi, "Automated condition assessment of concrete bridges with digital imaging," *Smart Structures and Systems*, vol. 13, no. 6, pp. 901–925, 2014. [Online]. Available: https://doi.org/10.12989/sss.2014.13.6.901.

[2] Y. Duan, Q. Chen, H. Zhang, C. B. Yun, S. Wu, and Q. Zhu, "CNN-based damage identification method of tied-arch bridge using spatial-spectral information," version 1.0, *Smart Structures and Systems*, vol. 23, no. 5, pp. 507–520, 2019. [Online]. Available: https://doi.org/10.12989/sss.2019.23.5.507 (visited on 10/27/2020).

[3] B. F. Spencer Jr, V. Hoskere, and Y. Narazaki, "Advances in computer vision-based civil infrastructure inspection and monitoring," *Engineering*, 2019. [Online]. Available: https://doi.org/10.1016/j.eng.2018.11.030.

[4] T. Yamaguchi and S. Hashimoto, "Fast crack detection method for large-size concrete surface images using percolation-based image processing," *Machine Vision and Applications*, vol. 21, no. 5, pp. 797–809, 2010. [Online]. Available: https://doi.org/10.1007/s00138-009-0189-8.

[5] H. Kim, E. Ahn, M. Shin, and S.-H. Sim, "Crack and noncrack classification from concrete surface images using machine learning," *Structural Health Monitoring*, vol. 18, no. 3, pp. 725–738, 2019. [Online]. Available: https://doi.org/10.1177/1475921718768747.

[6] F.-C. Chen and M. R. Jahanshahi, "NB-CNN: Deep learning-based crack detection using convolutional neural network and Naïve Bayes data fusion," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 5, pp. 4392–4400, 2017. [Online]. Available: https://doi.org/10.1109/TIE.2017.2764844.

[7] Y.-J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning-based crack damage detection using convolutional neural networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 5, pp. 361–378, 2017. [Online]. Available: https://doi.org/10.1111/mice.12263.

[8] M. R. Jahanshahi and S. F. Masri, "A new methodology for non-contact accurate crack width measurement through photogrammetry for automated structural safety evaluation," *Smart materials and structures*, vol. 22, no. 3, p. 035 019, 2013. [Online]. Available: https://doi.org/10.1088/0964-1726/22/3/035019.

[9] B. Shan, S. Zheng, and J. Ou, "A stereovision-based crack width detection approach for concrete surface assessment," *KSCE Journal of Civil Engineering*, vol. 20, no. 2, pp. 803–812, 2016. [Online]. Available: https://doi.org/10.1007/s12205-015-0461-6.

[10] M. R. Jahanshahi, F.-C. Chen, C. Joffe, and S. F. Masri, "Vision-based quantitative assessment of microcracks on reactor internal components of nuclear power plants," *Structure and Infrastructure Engineering*, vol. 13, no. 8, pp. 1013–1026, 2017. [Online]. Available: https://doi.org/10.1080/15732479.2016.1231207.

[11] S. German, I. Brilakis, and R. DesRoches, "Rapid entropy-based detection and properties measurement of concrete spalling with machine vision for post-earthquake safety assessments," *Advanced Engineering Informatics*, vol. 26, no. 4, pp. 846–858, 2012. [Online]. Available: https://doi.org/10.1016/j.aei.2012.06.005.

[12] I. Abdel-Qader, O. Abudayyeh, and M. E. Kelly, "Analysis of edge-detection techniques for crack identification in bridges," *Journal of Computing in Civil Engineering*, vol. 17, no. 4, pp. 255–263, 2003. [Online]. Available: https://doi.org/10.1061/(ASCE)0887-3801(2003)17:4(255).

[13] L. Ying and E. Salari, "Beamlet transform-based technique for pavement crack detection and classification," *Computer-Aided Civil and Infrastructure Engineering*, vol. 25, no. 8, pp. 572–580, 2010. [Online]. Available: https://doi.org/10.1111/j.1467-8667.2010.00674.x.

[14] C. Koch and I. Brilakis, "Pothole detection in asphalt pavement images," *Advanced Engineering Informatics*, vol. 25, no. 3, pp. 507–515, 2011. [Online]. Available: https://doi.org/10.1016/j.aei.2011.01.002.

[15] Q. Zou, Y. Cao, Q. Li, Q. Mao, and S. Wang, "Cracktree: Automatic crack detection from pavement images," *Pattern Recognition Letters*, vol. 33, no. 3, pp. 227–238, 2012. [Online]. Available: https://doi.org/10.1016/j.patrec.2011.11.004.

[16] E. Buza, S. Omanovic, and A. Huseinovic, "Pothole detection with image processing and spectral clustering," in *Proceedings of the 2nd International Conference on Information Technology and Computer Networks*, 2013, pp. 48–53.

[17] E. Zalama, J. Gómez-García-Bermejo, R. Medina, and J. Llamas, "Road crack detection using visual features extracted by gabor filters," *Computer-Aided Civil and Infrastructure Engineering*, vol. 29, no. 5, pp. 342–358, 2014. [Online]. Available: https://doi.org/10.1111/mice.12042.

[18] M. Avila, S. Begot, F. Duculty, and T. S. Nguyen, "2D image based road pavement crack detection by calculating minimal paths and dynamic programming," in *Image Processing (ICIP), 2014 IEEE International Conference on*, IEEE, 2014, pp. 783–787. [Online]. Available: https://doi.org/10.1109/ICIP.2014.7025157.

[19] R. Amhaz, S. Chambon, J. Idier, and V. Baltazart, "Automatic crack detection on two-dimensional pavement images: An algorithm based on minimal path selection.," *IEEE Trans. Intelligent Transportation Systems*, vol. 17, no. 10, pp. 2718–2729, 2016. [Online]. Available: https://doi.org/10.1109/TITS.2015.2477675.

[20] S. K. Sinha, P. W. Fieguth, and M. A. Polak, "Computer vision techniques for automatic structural assessment of underground pipes," *Computer-Aided Civil and Infrastructure Engineering*, vol. 18, no. 2, pp. 95–112, 2003. [Online]. Available: https://doi.org/10.1111/1467-8667.00302.

[21] W. Zhang, Z. Zhang, D. Qi, and Y. Liu, "Automatic crack detection and classification method for subway tunnel safety monitoring," *Sensors*, vol. 14, no. 10, pp. 19 307–19 328, 2014. [Online]. Available: https://doi.org/10.3390/s141019307.

[22] M. O'Byrne, F. Schoefs, B. Ghosh, and V. Pakrashi, "Texture analysis based damage detection of ageing infrastructural elements," *Computer-Aided Civil and Infrastructure Engineering*, vol. 28, no. 3, pp. 162–177, 2013. [Online]. Available: https://doi.org/10.1111/j.1467-8667.2012.00790.x.

[23] P.-H. Chen, H.-K. Shen, C.-Y. Lei, and L.-M. Chang, "Support-vector-machine-based method for automated steel bridge rust assessment," *Automation in Construction*, vol. 23, pp. 9–19, 2012. [Online]. Available: https://doi.org/10.1016/j.autcon.2011.12.001.

[24] A. Cord and S. Chambon, "Automatic road defect detection by textural pattern recognition based on adaboost," *Computer-Aided Civil and Infrastructure Engineering*, vol. 27, no. 4, pp. 244–259, 2012. [Online]. Available: https://doi.org/10.1111/j.1467-8667.2011.00736.x.

[25] Y. O. Ouma and M. Hahn, "Pothole detection on asphalt pavements from 2D-colour pothole images using fuzzy c-means clustering and morphological reconstruction," *Automation in Construction*, vol. 83, pp. 196–211, 2017. [Online]. Available: https://doi.org/10.1016/j.autcon.2017.08.017.

[26] L. Zhang, F. Yang, Y. D. Zhang, and Y. J. Zhu, "Road crack detection using deep convolutional neural network," in *Image Processing (ICIP), 2016 IEEE International Conference on*, IEEE, 2016, pp. 3708–3712. [Online]. Available: https://doi.org/10.1109/ICIP.2016.7533052.

[27] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, and A. Agrawal, "Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection," *Construction and Building Materials*, vol. 157, pp. 322–330, 2017. [Online]. Available: https://doi.org/10.1016/j.conbuildmat.2017.09.110.

[28] A. Zhang, K. C. Wang, B. Li, E. Yang, X. Dai, Y. Peng, Y. Fei, Y. Liu, J. Q. Li, and C. Chen, "Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 10, pp. 805–819, 2017. [Online]. Available: https://doi.org/10.1111/mice.12297.

[29] Y.-J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, and O. Büyüköztürk, "Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types," *Computer-Aided Civil and Infrastructure Engineering*, 2017. [Online]. Available: https://doi.org/10.1111/mice.12334.

[30] C. M. Yeum, S. J. Dyke, and J. Ramirez, "Visual data classification in post-event building reconnaissance," *Engineering Structures*, vol. 155, pp. 16–24, 2018. [Online]. Available: https://doi.org/10.1016/j.engstruct.2017.10.057.

[31] H. Kim, E. Ahn, M. Shin, and S.-H. Sim, "Crack and noncrack classification from concrete surface images using machine learning," *Structural Health Monitoring*, p. 1 475 921 718 768 747, 2018. [Online]. Available: https://doi.org/10.1177/1475921718768747.

[32] F.-C. Chen and M. R. Jahanshahi, "NB-CNN: Deep learning-based crack detection using convolutional neural network and naive bayes data fusion," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 5, pp. 4392–4400, 2018. [Online]. Available: https://doi.org/10.1109/TIE.2017.2764844.

[33] V. Hoskere, Y. Narazaki, T. Hoang, and B. Spencer Jr, "Vision-based structural inspection using multiscale deep convolutional neural networks," *arXiv preprint arXiv:1805.01055*, 2018.

[34] Y. Gao and K. M. Mosalam, "Deep transfer learning for image-based structural damage recognition," *Computer-Aided Civil and Infrastructure Engineering*, [Online]. Available: https://doi.org/10.1111/mice.12363.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826. [Online]. Available: https://doi.org/10.1109/CVPR.2016.308.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/cvpr.2016.90.

[37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning.," in *AAAI*, vol. 4, 2017, p. 12. [Online]. Available: https://dl.acm.org/doi/10.5555/3298023.3298188.

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

[39] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587. [Online]. Available: https://doi.org/10.1109/CVPR.2014.81.

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

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9. [Online]. Available: https://doi.org/10.1109/CVPR.2015.7298594.

[42] P. Shah, S. Pujol, A. Puranam, and L. Laughery, *Database on performance of low-rise reinforced concrete buildings in the 2015 nepal earthquake*, Jul. 2015. [Online]. Available: https://datacenterhub.org/resources/238.

[43] P. Shah, S. Pujol, and A. Puranam, *Database on performance of high-rise reinforced concrete buildings in the 2015 nepal earthquake*, Jul. 2015. [Online]. Available: https://datacenterhub.org/resources/242.

[44] C. Sim, E. Villalobos, J. P. Smith, P. Rojas, S. Pujol, A. Y. Puranam, and L. Laughery, *Performance of low-rise reinforced concrete buildings in the 2016 Ecuador earthquake*, Aug. 2016. [Online]. Available: https://datacenterhub.org/resources/14160.

[45] C. Sim, C. Song, N. Skok, A. Irfanoglu, S. Pujol, and M. Sozen, *Database of low-rise reinforced concrete buildings with earthquake damage*, Mar. 2015. [Online]. Available: https://datacenterhub.org/resources/123.

[46] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2015.

[47] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[48] J. H. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression.," in *CVPR*, 2017, pp. 6469–6477. [Online]. Available: https://doi.org/10.1109/CVPR.2017.685.

[49] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer, 2014, pp. 740–755. [Online]. Available: https://doi.org/10.1007/978-3-319-10602-1_48.

[50] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[51] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, Springer, 2010, pp. 177–186. [Online]. Available: https://doi.org/10.1007/978-3-7908-2604-3_16.

[52] R.-T. Wu, A. Singla, M. R. Jahanshahi, E. Bertino, B. J. Ko, and D. Verma, "Pruning deep convolutional neural networks for efficient edge computing in condition assessment of infrastructures," *Computer-Aided Civil and Infrastructure Engineering*, 2019. [Online]. Available: https://doi.org/10.1111/mice.12449.

[53] X. Kong and J. Li, "Non-contact fatigue crack detection in civil infrastructure through image overlapping and crack breathing sensing," *Automation in Construction*, vol. 99, pp. 125–139, 2019. [Online]. Available: https://doi.org/10.1016/j.autcon.2018.12.011.

[54] X. Kong and J. Li, "Vision-based fatigue crack detection of steel structures using video feature tracking," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 9, pp. 783–799, 2018. [Online]. Available: https://doi.org/10.1111/mice.12353.

[55] M. R. Jahanshahi, S. F. Masri, and G. S. Sukhatme, "Multi-image stitching and scene reconstruction for evaluating defect evolution in structures," *Structural Health Monitoring*, vol. 10, no. 6, pp. 643–657, 2011. [Online]. Available: https://doi.org/10.1177/1475921710395809.

[56] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European conference on computer vision*, Springer, 2006, pp. 404–417. [Online]. Available: https://doi.org/10.1007/11744023_32.

[57] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quarterly of applied mathematics*, vol. 2, no. 2, pp. 164–168, 1944. [Online]. Available: https://doi.org/10.1090/QAM/10666.

[58] D. W. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963. [Online]. Available: https://doi.org/10.1137/0111030.

[59] B. Y. Lee, Y. Y. Kim, S.-T. Yi, and J.-K. Kim, "Automated image processing technique for detecting and analysing concrete surface cracks," *Structure and Infrastructure Engineering*, vol. 9, no. 6, pp. 567–577, 2013. [Online]. Available: https://doi.org/10.1080/15732479.2011.593891.

[60] Z. Zhu, S. German, and I. Brilakis, "Visual retrieval of concrete crack properties for automated post-earthquake structural safety evaluation," *Automation in Construction*, vol. 20, no. 7, pp. 874–883, 2011. [Online]. Available: https://doi.org/10.1016/j.autcon.2011.03.004.

[61] B. Esser and D. R. Huston, "Versatile robotic platform for structural health monitoring and surveillance," *Smart Structures and Systems*, vol. 1, no. 4, pp. 325–338, 2005. [Online]. Available: https://doi.org/10.12989/sss.2005.1.4.325.

[62] C. Boller, P. Starke, G. Dobmann, C.-M. Kuo, and C.-H. Kuo, "Approaching the assessment of ageing bridge infrastructure," *Smart Structures and Systems*, vol. 15, no. 3, pp. 593–608, 2015. [Online]. Available: https://doi.org/10.12989/sss.2015.15.3.593.

[63] H. Myung, Y. Wang, S. Kang, and X. Chen, "Survey on robotics and automation technologies for civil infrastructure," 2014. [Online]. Available: https://doi.org/10.12989/sss.2014.13.6.891.

[64] C. M. Yeum, J. Choi, and S. J. Dyke, "Automated region-of-interest localization and classification for vision-based visual assessment of civil infrastructure," *Structural Health Monitoring*, vol. 18, no. 3, pp. 675–689, 2019. [Online]. Available: https://doi.org/10.1177/1475921718765419.

[65] C. Yeum, J. Choi, and S. Dyke, "Autonomous image localization for visual inspection of civil infrastructure," *Smart Materials and Structures*, vol. 26, no. 3, p. 035 051, 2017. [Online]. Available: https://doi.org/10.1088/1361-665X/aa510e.

[66] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94.

[67] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981. [Online]. Available: https://doi.org/10.1145/358669.358692.

[68] M. R. Jahanshahi, J. S. Kelly, S. F. Masri, and G. S. Sukhatme, "A survey and evaluation of promising approaches for automatic image-based defect detection of bridge structures," *Structure and Infrastructure Engineering*, vol. 5, no. 6, pp. 455–486, 2009. [Online]. Available: https://doi.org/10.1080/15732470801945930.

[69] Y.-J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, and O. Büyüköztürk, "Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 9, pp. 731–747, 2018. [Online]. Available: https://doi.org/10.1111/mice.12334.

[70] H. Minkowski, "Volumen und oberflache, volume 447," *Math. Annalen*, vol. 57, 1903. [Online]. Available: https://doi.org/10.1090/S0002-9904-1975-13853-5.

[71] G. Matheron, "Random sets and integral geometry," 1975. [Online]. Available: https://doi.org/10.1090/S0002-9904-1975-13853-5.

[72] P. Salembier, "Comparison of some morphological segmentation algorithms based on contrast enhancement. application to automatic defect detection.," in *5. European Signal Processing Conference.*, vol. 2, 1990, pp. 833–836.

[73] M. R. Jahanshahi, S. F. Masri, C. W. Padgett, and G. S. Sukhatme, "An innovative methodology for detection and quantification of cracks through incorporation of depth perception," *Machine vision and applications*, vol. 24, no. 2, pp. 227–241, 2013. [Online]. Available: https://doi.org/10.1007/s00138-011-0394-0.

[74] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979. [Online]. Available: http://dx.doi.org/10.1109/TSMC.1979.4310076.

[75] S.-N. Yu, J.-H. Jang, and C.-S. Han, "Auto inspection system using a mobile robot for detecting concrete cracks in a tunnel," *Automation in Construction*, vol. 16, no. 3, pp. 255–261, 2007. [Online]. Available: https://doi.org/10.1016/j.autcon.2006.05.003.

[76] R. Van Uitert and I. Bitter, "Subvoxel precise skeletons of volumetric data based on fast marching methods," *Medical physics*, vol. 34, no. 2, pp. 627–638, 2007. [Online]. Available: https://doi.org/10.1118/1.2409238.

[77] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke, "RGB-D object detection and semantic segmentation for autonomous manipulation in clutter," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 437–451, 2018. [Online]. Available: https://doi.org/10.1177/0278364917713117.

[78] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Asian conference on computer vision*, Springer, 2016, pp. 213–228. [Online]. Available: https://doi.org/10.1007/978-3-319-54181-5_14.

[79] Z. Li, Y. Gan, X. Liang, Y. Yu, H. Cheng, and L. Lin, "LSTM-CF: Unifying context modeling and fusion with LSTMs for RGB-D scene labeling," in *European conference on computer vision*, Springer, 2016, pp. 541–557. [Online]. Available: https://doi.org/10.1007/978-3-319-46475-6_34.

[80] J. Wang, Z. Wang, D. Tao, S. See, and G. Wang, "Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks," in *European Conference on Computer Vision*, Springer, 2016, pp. 664–679. [Online]. Available: https://doi.org/10.1007/978-3-319-46454-1_40.

[81] S.-J. Park, K.-S. Hong, and S. Lee, "RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4980–4989. [Online]. Available: https://doi.org/10.1109/ICCV.2017.533.

[82] Y. Cheng, R. Cai, Z. Li, X. Zhao, and K. Huang, "Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3029–3037. [Online]. Available: https://doi.org/10.1109/CVPR.2017.161.

[83] X. Xu, Y. Li, G. Wu, and J. Luo, "Multi-modal deep feature learning for RGB-D object detection," *Pattern Recognition*, vol. 72, pp. 300–313, 2017. [Online]. Available: https://doi.org/10.1016/j.patcog.2017.07.026.

[84] T. Ophoff, K. Van Beeck, and T. Goedemé, "Improving real-time pedestrian detectors with RGB+ depth fusion," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, IEEE, 2018, pp. 1–6. [Online]. Available: https://doi.org/10.1109/AVSS.2018.8639110.

[85] K. Zhou, A. Paiement, and M. Mirmehdi, "Detecting humans in rgb-d data with cnns," in *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*, IEEE, 2017, pp. 306–309. [Online]. Available: https://doi.org/10.23919/MVA.2017.7986862.

[86] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng, "Convolutional-recursive deep learning for 3D object classification," in *Advances in neural information processing systems*, 2012, pp. 656–664.

[87] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2015, pp. 681–687. [Online]. Available: https://doi.org/10.1109/IROS.2015.7353446.

[88] A. Wang, J. Lu, J. Cai, T.-J. Cham, and G. Wang, "Large-margin multi-modal deep learning for RGB-D object recognition," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1887–1898, 2015. [Online]. Available: https://doi.org/10.1109/TMM.2015.2476655.

[89] Y. Gao and K. M. Mosalam, "Deep transfer learning for image-based structural damage recognition," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 9, pp. 748–768, 2018. [Online]. Available: https://doi.org/10.1111/mice.12363.

[90] S. Zhou and W. Song, "Deep learning–based roadway crack classification with heterogeneous image data fusion," *Structural Health Monitoring*, vol. 20, no. 3, pp. 1274–1293, 2021. [Online]. Available: https://doi.org/10.1177/1475921720948434.

[91] S. Zhou and W. Song, "Deep learning-based roadway crack classification using laser-scanned range images: A comparative study on hyperparameter selection," *Automation in Construction*, vol. 114, p. 103 171, 2020. [Online]. Available: https://doi.org/10.1016/j.autcon.2020.103171.

[92] S. Zhou and W. Song, "Concrete roadway crack segmentation using encoder-decoder networks with range images," *Automation in Construction*, vol. 120, p. 103 403, 2020. [Online]. Available: https://doi.org/10.1016/j.autcon.2020.103403.

[93] H. Kim, S. Lee, E. Ahn, M. Shin, and S.-H. Sim, "Crack identification method for concrete structures considering angle of view using RGB-D camera-based sensor fusion," *Structural Health Monitoring*, p. 1 475 921 720 934 758, 2020. [Online]. Available: https://doi.org/10.1177/1475921720934758.

[94] G. H. Beckman, D. Polyzois, and Y.-J. Cha, "Deep learning-based automatic volumetric damage quantification using depth camera," *Automation in Construction*, vol. 99, pp. 114–124, 2019. [Online]. Available: https://doi.org/10.1016/j.autcon.2018.12.006.

[95] M. R. Jahanshahi, F. Jazizadeh, S. F. Masri, and B. Becerik-Gerber, "Unsupervised approach for autonomous pavement-defect detection and quantification using an inexpensive depth sensor," *Journal of Computing in Civil Engineering*, vol. 27, no. 6, pp. 743–754, 2013. [Online]. Available: https://doi.org/10.1061/(ASCE)CP.1943-5487.0000245.

[96] H. Indie. [Online]. Available: https://www.sidefx.com/products/houdini/houdini-indie/.

[97] A. A. Standard, "Building code requirements for structural concrete (ACI 318-11)," in *American Concrete Institute*, 2011.

[98] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3234–3243. [Online]. Available: https://doi.org/10.1109/CVPR.2016.352.

[99] V. Hoskere, Y. Narazaki, B. F. Spencer, and M. D. Smith, "Deep learning-based damage detection of miter gates using synthetic imagery from computer graphics," in *12th International Workshop on Structural Health Monitoring: Enabling Intelligent Life-Cycle Health Management for Industry Internet of Things (IIOT), IWSHM 2019*, DEStech Publications Inc., 2019, pp. 3073–3080. [Online]. Available: https://doi.org/10.12783/shm2019/32463.

[100] V. Hoskere, Y. Narazaki, and B. Spencer, "Learning to detect important visual changes for structural inspections using physicsbased graphics models," in *9th International Conference on Structural Health Monitoring of Intelligent Infrastructure: Transferring Research into Practice, SHMII 2019*, International Society for Structural Health Monitoring of Intelligent …, 2019, pp. 1484–1490.

[101] Y. Narazaki, V. Hoskere, K. Yoshida, B. F. Spencer, and Y. Fujino, "Synthetic environments for vision-based structural condition assessment of japanese high-speed railway viaducts," *Mechanical Systems and Signal Processing*, vol. 160, p. 107 850, 2021. [Online]. Available: https://doi.org/10.1016/j.ymssp.2021.107850.

[102] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 826–834.

[103] K. Gunasekar, Q. Qiu, and Y. Yang, "Low to high dimensional modality hallucination using aggregated fields of view," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1983–1990, 2020. [Online]. Available: https://doi.org/10.1109/LRA.2020.2970679.

[104] A. Bhoi, "Monocular depth estimation: A survey," *arXiv preprint arXiv:1901.09402*, 2019.

[105] C. Zhao, Q. Sun, C. Zhang, Y. Tang, and F. Qian, "Monocular depth estimation based on deep learning: An overview," *Science China Technological Sciences*, pp. 1–16, 2020. [Online]. Available: https://doi.org/10.1007/s11431-020-1582-8.

[106] D. Dornia and N. Fischöder, "Sidefx houdini in vfx," in *Interaktive Datenvisualisierung in Wissenschaft und Unternehmenspraxis*, Springer, 2020, pp. 21–44. [Online]. Available: https://doi.org/10.1007/978-3-658-29562-2_2.

[107] A. Abgottspon, "Procedural modelling in Houdini based on function representation,"

[108] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from RGB-D data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017. [Online]. Available: https://doi.org/10.1109/3DV.2017.00081.

[109] K. Perlin, "Noise hardware. in real-time shading'," *SIGGRAPH Course Notes*, 2001.

[110] G. R. Hofmann, "Who invented ray tracing?" *The Visual Computer*, vol. 6, no. 3, pp. 120–124, 1990. [Online]. Available: https://doi.org/10.1007/BF01911003.

[111] M. Pharr, W. Jakob, and G. Humphreys, *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016. [Online]. Available: https://doi.org/10.1016/C2013-0-15557-2.

[112] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[113] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 2009, pp. 248–255. [Online]. Available: https://doi.org/10.1109/CVPR.2009.5206848.

[114] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *2011 International Conference on Computer Vision*, IEEE, 2011, pp. 2018–2025. [Online]. Available: https://doi.org/10.1109/ICCV.2011.6126474.

[115] S. Zennaro, M. Munaro, S. Milani, P. Zanuttigh, A. Bernardi, S. Ghidoni, and E. Menegatti, "Performance evaluation of the 1st and 2nd generation kinect for multimedia applications," in *2015 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2015, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICME.2015.7177380.

[116] A. G. Taylor, "What is the microsoft hololens?" In *Develop Microsoft HoloLens Apps Now*, Springer, 2016, pp. 3–7. [Online]. Available: https://doi.org/10.1007/978-1-4842-2202-7_1.

[117] O. Özyeşil, V. Voroninski, R. Basri, and A. Singer, "A survey of structure from motion*.," *Acta Numerica*, vol. 26, pp. 305–364, 2017. [Online]. Available: https://doi.org/10.1017/S096249291700006X.

[118] S. Ullman, "The interpretation of structure from motion," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 203, no. 1153, pp. 405–426, 1979. [Online]. Available: https://doi.org/10.1098/rspb.1979.0006.

[119] C. Wu *et al.*, "VisualSFM: A visual structure from motion system," 2011.

[120] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[121] Z.-L. Cao, Z.-H. Yan, and H. Wang, "Summary of binocular stereo vision matching technology," *Journal of Chongqing University of Technology (Natural Science)*, vol. 29, no. 2, pp. 70–75, 2015.

[122] L. Zou and Y. Li, "A method of stereo vision matching based on OpenCV," in *2010 International Conference on Audio, Language and Image Processing*, IEEE, 2010, pp. 185–190. [Online]. Available: https://doi.org/10.1109/ICALIP.2010.5684978.

[123] N. Lazaros, G. C. Sirakoulis, and A. Gasteratos, "Review of stereo vision algorithms: From software to hardware," *International Journal of Optomechatronics*, vol. 2, no. 4, pp. 435–462, 2008. [Online]. Available: https://doi.org/10.1080/15599610802438680.

[124] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[125] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*, 2018.

[126] R. Groenendijk, S. Karaoglu, T. Gevers, and T. Mensink, "On the benefit of adversarial training for monocular depth estimation," *Computer Vision and Image Understanding*, vol. 190, p. 102 848, 2020. [Online]. Available: https://doi.org/10.1016/j.cviu.2019.102848.

[127] D.-H. Kwak and S.-H. Lee, "A novel method for estimating monocular depth using cycle gan and segmentation," *Sensors*, vol. 20, no. 9, p. 2567, 2020. [Online]. Available: https://doi.org/10.3390/s20092567.

[128] K. Gwn Lore, K. Reddy, M. Giering, and E. A. Bernal, "Generative adversarial networks for depth map estimation from RGB video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1177–1185. [Online]. Available: https://doi.org/10.1109/CVPRW.2018.00163.

[129] A. CS Kumar, S. M. Bhandarkar, and M. Prasad, "Monocular depth prediction using generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 300–308. [Online]. Available: https://doi.org/10.1109/CVPRW.2018.00068.

[130] D. S. Tan, C.-Y. Yao, C. Ruiz, and K.-L. Hua, "Single-image depth inference using generative adversarial networks," *Sensors*, vol. 19, no. 7, p. 1708, 2019. [Online]. Available: https://doi.org/10.3390/s19071708.

[131] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[132] T. A. Cruse and P. Besuner, "Residual life prediction for surface cracks in complex structural details," *Journal of Aircraft*, vol. 12, no. 4, pp. 369–375, 1975. [Online]. Available: https://doi.org/10.2514/3.44458.

[133] S. K. Dhawan, A. Bindal, S. Bhalla, and B. Bhattacharjee, "Expected residual service life of reinforced concrete structures from current strength considerations," *Advances in Structural Engineering*, vol. 22, no. 7, pp. 1631–1643, 2019. [Online]. Available: https://doi.org/10.1177/1369433218818001.

[134] P. Besuner, "Residual life estimates for structures with partial thickness cracks," in *Mechanics of crack growth*, ASTM International, 1976. [Online]. Available: https://doi.org/10.1520/STP33959S.

[135] X. Wang, N. Balakrishnan, and B. Guo, "Residual life estimation based on a generalized wiener degradation process," *Reliability Engineering & System Safety*, vol. 124, pp. 13–23, 2014. [Online]. Available: https://doi.org/10.1016/j.ress.2013.11.011.

[136] A. K. Agrawal, A. Kawaguchi, and Z. Chen, "Deterioration rates of typical bridge elements in new york," *Journal of Bridge Engineering*, vol. 15, no. 4, pp. 419–429, 2010. [Online]. Available: https://doi.org/10.1061/(ASCE)BE.1943-5592.0000123.

[137] S. Mohanty, A. Chattopadhyay, and P. Peralta, "Adaptive residual useful life estimation of a structural hotspot," *Journal of Intelligent Material Systems and Structures*, vol. 21, no. 3, pp. 321–335, 2010. [Online]. Available: https://doi.org/10.1177/1045389X09357972.

[138] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.