Semantic Segmentation of Sewer Pipe Defects Using Deep Dilated Convolutional Neural Network

M.Z. Wang[a] and J.C.P. Cheng[a]

[a] Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
E-mail: mwangaz@connect.ust.hk, cejcheng@ust.hk

Abstract –

Semantic segmentation of closed-circuit television (CCTV) images can facilitate automatic severity assessment of sewer pipe defects by assigning defect labels to each pixel in the image, from which defect types, locations and geometric information can be obtained. In this study, a deep convolutional neural network (CNN), namely DilaSeg, is developed based on dilated convolution for improving the segmentation of sewer pipe defects including cracks, tree root intrusion and deposit. Sewer pipe CCTV images are extracted from inspection videos and are annotated to be used as the ground truth labels for training the model. DilaSeg is constructed with dilated convolution for producing feature maps with high resolution. Both DilaSeg and the state-of-the-art model, fully convolutional network (FCN), are trained and evaluated on the annotated dataset using the same hyper-parameters. The results of the experiments indicate that the proposed DilaSeg improved the segmentation accuracy significantly compared with FCN, with 18% of increase in mean pixel accuracy (mPA) and 22% of increase in mean intersection over union (IoU) with a fast detection speed.

Keywords –

Dilated convolution; Convolutional neural network (CNN); Semantic segmentation; Sewer pipe defect; Defect segmentation; Visual inspection

1 Introduction

Sewer pipe defects such as cracks, root intrusions through the pipe joints and the deposits inside the pipe are major causes of pipe deterioration, leading to serious consequences e.g. around 23,000 to 75,000 sanitary sewer overflows (SSOs) occurred every year in the United States, causing concerns for the environment and human life [1]. Therefore, discovering and repairing pipe defects at an early stage is significant to prevent sewer system deterioration such that severe consequences can be avoided. Currently, visual inspection techniques such as closed-circuit television (CCTV) are widely utilized for sewer pipe inspection. As the manual assessment is time-consuming, error-prone and subjective, computer vision techniques are studied for defect classification, defect detection and semantic segmentation. Defect detection and classification can inform inspectors of defect type and relative locations in the image. However, automated assessment of defect severity has been rarely studied, although it is important for arranging maintenance activities. Semantic segmentation can facilitate defect severity assessment by providing the defect type, location and geometric information for each pixel.

Semantic segmentation traditionally relies on classifiers such as Support Vector Machines (SVMs) or probabilistic graphical models e.g. Markov Random Fields (MRFs) and Conditional Random Fields (CRFs) for modelling pixel relationships [2], which all require handcrafted feature descriptors. In recent years, deep learning based models such as fully convolutional network (FCN) are developed by learning rich features automatically and have achieved the state-of-the-art performance [3]. However, the consecutive convolution and pooling implementations in previous deep learning models down-sampled the feature maps by a large factor and generate feature maps of low resolution. Such severe spatial information loss results in obscure segmentation for detailed structures or object boundaries after up-sampling. Therefore, the objective of this study is to improve sewer pipe defect segmentation by proposing a deep CNN model, namely DilaSeg, based on normal convolution, dilated convolution and bilinear interpolation. The proposed model is constructed to extract important features by normal convolution, improve the feature map resolution by dilated convolution, refine the segmentation for object with multiple scales by multi-scale dilated convolution and upsample the feature maps using bilinear interpolation. Sewer pipe images with three types of defects are collected and annotated to train both the proposed model and FCN. Model performance is evaluated and analyzed through experiments.

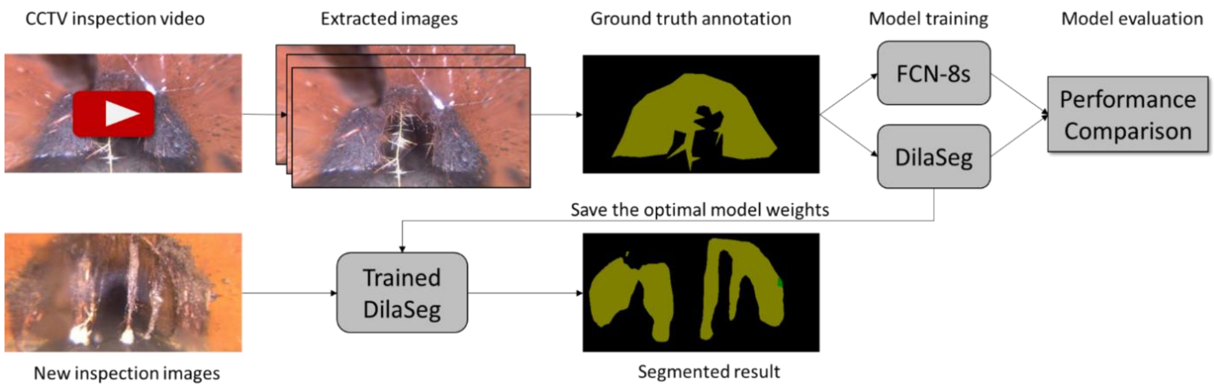


Figure 1. The workflow of sewer pipe defect segmentation using deep learning models

2 Related Work

Computer vision techniques are studied to facilitate the automatic interpretation of visual sewer inspection results. Previous studies have focused on the classification and detection of sewer pipe defects from CCTV images using image processing techniques e.g. morphological operations, histograms of oriented gradients (HOG) and SVMs [4]. More recently, deep learning is applied to address limitations of conventional methods by learning rich features automatically with convolutional neural networks (CNNs). A CNN based model was proposed for classifying multiple sewer pipe defects from CCTV images [5] and a region-based CNN method has been proposed for identifying and locating defects with bounding boxes on sewer pipe images [6]. So far, segmentation of pipe images mainly focuses on pipe components such as joints or segmenting single types of defects using image processing techniques [7]. There are limited studies on the segmentation of multiple pipe defects of sewer pipes, which could provide important references for evaluating the defect severity.

In conventional semantic segmentation, feature extractors such as HOG or scale-invariant feature transform (SIFT) are designed to extract expressive features, based on which small patches of the input image are extracted and classified using local classifiers like random decision forest or SVMs [8]. However, small spatial windows often lead to noisy prediction and require high computational cost. MRFs and CRFs have been widely applied for semantic segmentation [9] by modelling the correlations between variables and incorporating the context knowledge. The main limitation of MRFs and CRFs is that features are obtained from conventional classifiers which are designed for particular cases, during which expertise and efforts are needed. On the contrast, FCN is one breakthrough semantic segmentation deep learning model with higher accuracy than traditional approaches. FCN is developed by transforming the fully connected layers of typical CNNs into convolutional layers and developing a deconvolutional layer for upsampling feature maps [3]. Based on the FCN, many variant networks are proposed, utilizing the architecture of a typical CNN as the “encoder” for generating feature maps and focusing on the development of “decoder” for upsampling images [10]. However, the feature maps generated by FCN and other similar models are of low resolution and predictions are coarse due to the huge spatial information loss during downsampling process.

3 Methodology

3.1 Workflow of Sewer Pipe Defect Segmentation Using Deep Learning Models

As shown in Figure 1, the workflow of implementing deep learning models for segmenting sewer pipe defects mainly includes: (1) extract image from CCTV inspection videos; (2) pre-process and annotate images with different colors for each defect as ground truth labels; (3) build the model architecture and train the model using the annotated images; (4) evaluate and compare the proposed model with the state-of-the-art model; (5) save the best model and apply for new images. Among all the steps in the workflow, the model architecture has great influence on the segmentation performance while the model evaluation provides metrics on the performance. Details of the model and the evaluation are introduced in the following sections.

3.1.1 Architecture of the Proposed DilaSeg

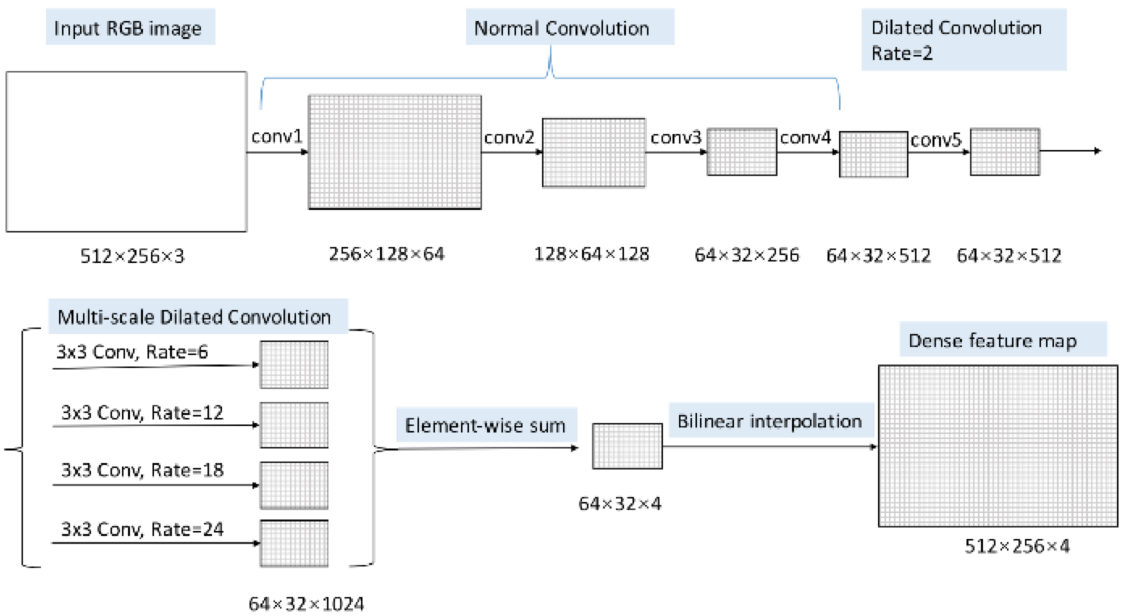


Figure 2. Brief architecture of DilaSeg for generating dense feature maps

To address the problem of large spatial information loss in previous deep learning models, DilaSeg is proposed to improve the feature map resolution and to produce dense feature maps. As shown in Figure 2, the dense feature maps are obtained mainly through (1) normal convolution, (2) dilated convolution, (3) multi-scale dilated convolution and (4) bilinear interpolation. After obtaining the dense feature maps, softmax function is applied for calculating the loss and producing the predicted labels for each pixel.

(1) Normal convolution

During the normal convolution, each image is fed into the network as a three-channel array and a certain number of filters are assigned with random weights at the model initialization stage. In the convolutional layer, the dot product of filters and the convolved image patch is calculated and added with a bias value get the convolution result. In the activation layer, the convolution result is fed into an activation function called Rectified Linear Units (ReLU) so as to add non-linearity to the model. The max-pooling layer is applied after the activation layer, during which only the maximum value in the covered feature map patch is remained to the next layer. The function of the max-pooling is to reduce the spatial dimension of the feature maps such that the computational cost will not exceed the capability. The process of performing a stack of convolution, ReLU and max pooling layers consecutively can be regarded as a down-sampling process, after which the obtained feature maps are of low resolution and may lead to difficulties in the later map upsampling process. Therefore, the resolution of feature maps is controlled through setting different zero padding layers around the feature maps in this study. In the end, the original images are only downsampled by a factor of 8, which is smaller than that by using other networks such as FCN and hence prevents large spatial information loss.

(2) Dilated convolution

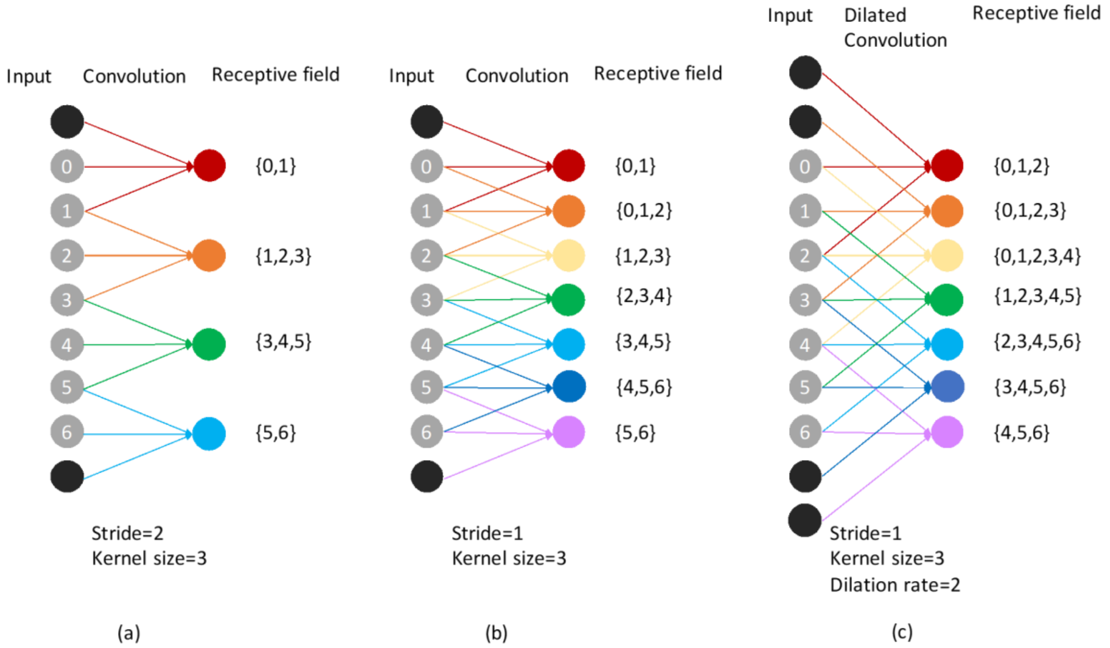


Figure 3. Example of normal convolution and dilated convolution

Another method to prevent too much resolution loss of feature maps is using smaller stride values during the convolution process. For example, the feature map resolution is increased in Figure 3 (b) compared with Figure 3 (a) as the stride value is decreased from 2 to 1. Nevertheless, one limitation of applying small stride values is that the receptive field (i.e. the area convolved by the filters with respect to the original input) is still small, which means there is not much global information incorporated during the learning process. Therefore, dilated convolution [11], is applied in the proposed model such that feature map resolution is improved and the receptive field is enlarged, but the parameter number will not increase. The dilated convolution filters can be treated as normal filters filled with certain number of holes (zeros), which is indicated by the dilation rate. Filters with dilation rate of k means that there are $(k\text{-}1)$ zeros inserted among each consecutive columns or rows of original filters. Specially, filters with dilation rate of 1 are the same with normal filters. Consequently, the scope of the convolved pixels of the original input (i.e. receptive field) is increased while the number of parameters remains the same with that of using normal filters. For example, Figure 3 (c) with the dilation rate of 2 increases the map resolution to the same density as (b) while the receptive field becomes larger, which enables more contextual information involved.

(3) Multi-scale dilated convolution

During segmentation, when aspect ratio of objects is different, convolution using the fix-sized filters only extract features from objects with certain scale while omitting features of different scale objects. Similar case exits for the segmentation of sewer pipe defects, considering different scales of the sewer pipe defects, e.g. scale of cracks is similar to long and thin rectangle while deposits tend to have square scale. Therefore, multi-scale dilated convolution layers are added in the developed model. Specifically, dilated convolution with 4 different dilation rates i.e. {6,12,18,24} is performed in parallel to generate feature maps respectively, the process of which is similar to convolution using filters of different sizes. With the aim not to reduce the feature map resolution, zero padding layers are applied according to the dilation rate of each layer. The feature map values from the four parallel layers are summed pixel-by-pixel and the fused results are used for the final interpolation.

(4) Bilinear interpolation

After the parallel dilated convolution, the generated feature maps are upsampled using bilinear interpolation, which is a common method used for upsampling. As shown in Figure 4, on the feature map to be upsampled, the value of each pixel is known. With the information of both location and values two pixels, the value of pixels in a certain location between them can be calculated through linear interpolation. Similarly, the value of pixels at other positions can also be obtained. For example, based on the known coordinates of all the pixels on the feature map, the value of pixel e can be obtained through the P_a and P_b while P_f can be calculated from P_c and P_d. In the end, P_g is obtained using P_e and P_f. By repeating this process, the feature map can be upsampled to the original resolution with values for each pixel.

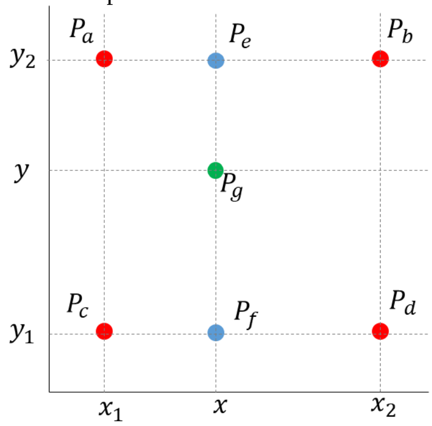


Figure 4. Example of bilinear interpolation

3.2 Model Evaluation

The proposed model is designed for real-time defect segmentation, the model implementation and computational cost should be feasible for on-site inspection while the segmentation accuracy should also be satisfying.

3.2.1 Accuracy

The four indices in [12] are used for measuring segmentation accuracy. As shown in Figure 5, assume the total class number is $\mathrm{k}+1$ (including k objects and 1 background class), $\mathrm{p_{ij}}$ is the number of pixels of class i but assigned with class j and can be treated as false positives. $\mathrm{p_{ji}}$ can be interpreted as false negatives while $\mathrm{p_{ii}}$ represents true positives. Circle A represents the total number of pixels of class i, i.e. $\mathrm{t_i} = \sum_{\mathrm{j}=0}^{\mathrm{k}} \mathrm{p_{ij}}$. Circle B represents all the pixels predicted to be class i and the intersection C between A and B represents all the true positives.

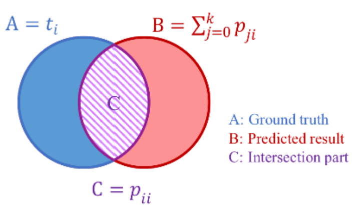


Figure 5. Graphical representation of the predicted results and ground truth

(1) Pixel accuracy (PA)

PA is the simplest evaluation by calculating the ratio of pixels correctly classified over the total number of pixels.

$$\mathrm{PA} = \frac{\sum_{\mathrm{i}=0}^{\mathrm{k}} \mathrm{p_{ii}}}{\sum_{\mathrm{i}=0}^{\mathrm{k}} \mathrm{t_i}} \tag{8}$$

(2) Mean pixel accuracy (mPA)

mPA is calculated based on the pixel accuracy for each class, by computing the ratio of correctly predicted pixels over the total pixel amount in each class and taking the average value for all the classes.

$$\mathrm{mPA} = \frac{1}{\mathrm{k}+1} \sum_{\mathrm{i}=0}^{\mathrm{k}} \frac{\mathrm{p_{ii}}}{\mathrm{t_i}} \tag{9}$$

(3) Mean intersection over union (mIoU)

IoU is the dominant metric for evaluating segmentation accuracy and is calculated by taking the ratio of the intersection between predicted results and ground truth labels over the union between these two sets. The intersection is the true positives while the union is the sum of false positives and false negatives and subtracted by true positives. mIoU is obtained by taking average value of the IoUs for all the classes.

$$\mathrm{mIoU} = \frac{1}{\mathrm{k}+1} \sum_{\mathrm{i}=0}^{\mathrm{k}} \frac{\mathrm{p_{ii}}}{\mathrm{t_i} + \sum_{\mathrm{j}=0}^{\mathrm{k}} \mathrm{p_{ji}} - \mathrm{p_{ii}}} \tag{10}$$

(4) Frequency weighted intersection over union (fwIoU)

As the percentage of pixels belonging to each class may be different in the training dataset, evaluating the accuracy considering pixel occurrence frequency is also important.

$$\mathrm{fwIoU} = \frac{1}{\sum_{\mathrm{i}=0}^{\mathrm{k}} \sum_{\mathrm{j}=0}^{\mathrm{k}} \mathrm{p_{ij}}} \sum_{\mathrm{i}=0}^{\mathrm{k}} \frac{\mathrm{p_{ii}} \mathrm{t_i}}{\mathrm{t_i} + \sum_{\mathrm{j}=0}^{\mathrm{k}} \mathrm{p_{ji}} - \mathrm{p_{ii}}} \tag{11}$$

3.2.2 Computational Cost

The computational cost is a significant evaluation aspect to ensure a desired segmentation speed for real-time segmentation of sewer pipe defects. The computational cost of the models is evaluated for the inference stage using the segmentation speed as well as the training stage using training duration and model convergence performance. Compared with training stage, evaluation of the inference stage is more important as the inference speed affects the real-time performance of the model.

4 Experiments and Results

To validate the proposed models for sewer defect segmentation, experiments are conducted to evaluate

FCN-8s and the DilaSeg in terms of accuracy and computational cost.

4.1 Experiment Dataset and Implementation Details

Images containing three types of defects i.e. cracks, deposits and tree root intrusions are extracted from CCTV inspection videos of sewer pipe inspection company in the United States. 90% of 1510 annotated images are used for training (1,359 images) and 10% are for testing (151 images). The annotation files are generated through LabelMe [13] by plotting the polygon along the boundary of each defect. All the pixels inside the same polygon are assigned with the same defect class. In the end, each defect is annotated with different colors and the background is set to be black. Caffe [14] is one common library for building deep learning models. As the functions of the dilated convolution are not included in the initial Caffe, source code of Caffe was revised and recompiled for training the proposed model. All the segmentation models are trained on Ubuntu system with Intel® Core™ i7-6700 CPU @ 3.40GHz × 8 and GPU of GeForce GTX 1080. During each training iteration, images are fed into the network with a mini-batch of 16, to reduce GPU memory requirement and improve the training efficiency. The "poly" learning rate is applied with a base learning rate of 0.01, momentum of 0.9 and weight decay of 0.005. Each model is trained using the re-compiled Caffe for 50,000 iterations and saved every 500 iterations to evaluate their accuracy on validation dataset.

4.2 Experiment Results

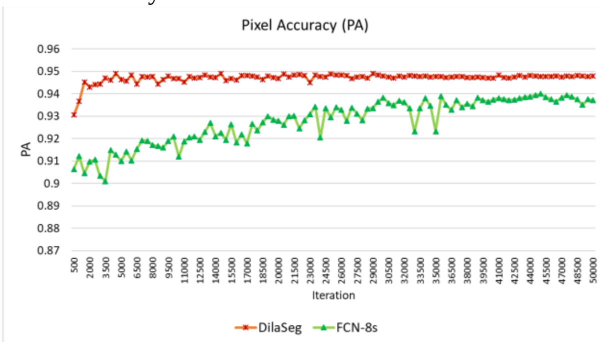
4.2.1 Accuracy

As shown in Figure 6 (a), the pixel accuracies of the two models follow a similar trend, i.e. within a certain number of iterations, PA increases with the increase of training iteration. The increase is obvious during first few thousand iterations and gradually becomes small in the later training period. Finally, the PA reaches a plateau and no longer yields increase. The point where the model reaching the plateau indicates the convergence of the model and is different for each model depending on the convergence speed. The DilaSeg has higher PA values than FCN-8s, indicating the effectiveness of dilated convolution.

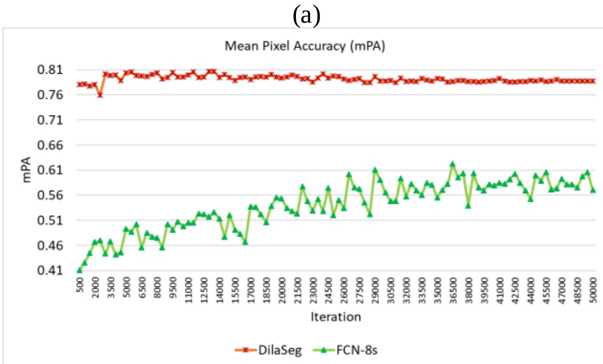
As shown in Figure 6 (b), the mPA values of both models are increasing during the training process. Although FCN-8s has a more obvious increase trend, DilaSeg achieved much higher mPA value than FCN-8s. As shown in Figure 6 (c), in terms of values of mIoU, the overall varying situation is similar to PA and mPA, with an increase at first and reaching a plateau in the end. The mIoU values of FCN-8s also increase with the training iterations, but the overall mIoU values are much lower than the proposed model. The mIoU values of DilaSeg are higher than FCN-8s during the whole training process, which reflecting the stronger capability of the developed model for segmentation. The fwIoU values which considers the pixel occurrence frequency when calculating IoU are shown in Figure 6 (d). Almost during the whole training, the fwIoU values of DilaSeg are higher than FCN-8s.

The best accuracy values of the two models are shown in Table 1. It can be seen that accuracies in terms of the four indices obtained using DilaSeg are higher than FCN-8s. The developed model improved the segmentation accuracy largely, especially with 18% of increase in mPA and 22% of increase in mIoU using DilaSeg.

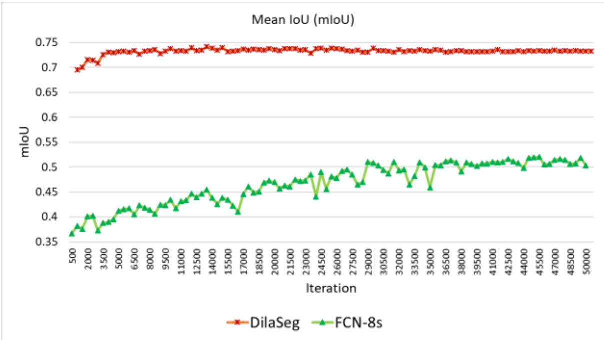
Some of the segmentation results are analysed. As shown in Figure 7 (a) and (b), for crack segmentation, FCN-8s can approximately provide the locations but cannot detect the correct defect type. One reason is that the model cannot recognize the defect type due to the color similarity between the cracks and other defects.



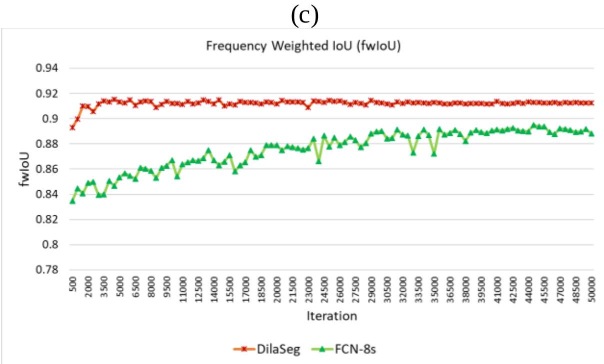
(a)



(b)



(c)



(d)

Figure 6. Accuracy of the two models for segmentation of sewer pipe defects

The DilaSeg performed better than FCN-8s, with segmentation of more parts of cracks, although some cracks are not segmented completely. In terms of deposit and tree root, FCN-8s achieved better results compared with segmenting cracks. However, segmentation of deposit and tree root are mixed with each other as shown in Figure 7 (c) and some defects are not segmented completely as shown in Figure 7 (d). On the contrary, DilaSeg is capable of segmenting both deposit and tree root much better with fewer mixed segmentation cases. In addition, more areas of the defects can be segmented by DilaSeg and more complete segmentation results can be provided as shwon in Figure 7 (e).

Table 1. Accuracy of the two models

	FCN-8s	DilaSeg
PA	0.939	0.949
mPA	0.623	0.807
mIoU	0.521	0.742
fwIoU	0.895	0.915

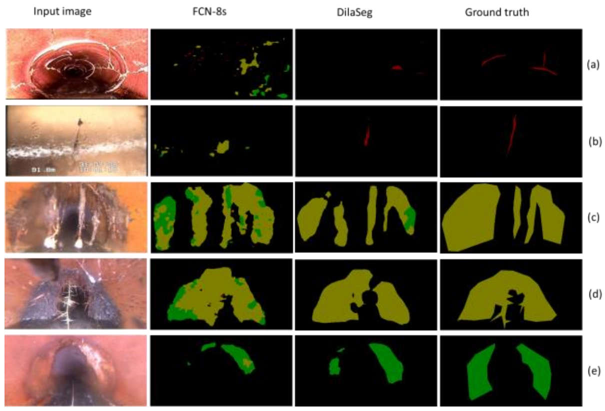


Figure 7. Examples of segmentation results (red color represents cracks, lime color represents roots and green color represents deposit)

4.2.2 Computational Cost

The computational cost of the two methods is evaluated using the detection speed (i.e. the time taken for detecting each image), training duration (in hours) and converge iteration (i.e. the number of iterations taken for the model to obtain convergence). The training duration and converge iteration indicate the difficulty for training the model to achieve desired performance while the detection speed reflects the possibility of real-time segmentation. As shown in Table 2, although the training process of DilaSeg is longer than FCN-8s, the detection speed of DilaSeg is relatively faster during model inference. In addition, the training loss of DilaSeg is dropping quickly within the first few thousand iterations and achieved a plateau at around 15,000 iterations, as shown in Figure 8 (a). However, the loss of FCN-8s was much higher and was dropping at a quite slow rate during the whole process, achieving a converging point after 45,000 iterations. The training loss trend reflects the proposed model can learn image features and optimize weights more efficiently.

Table 2. Computational cost of the two models

	FCN-8s	DilaSeg
Detection speed (s/image)	0.352	0.265
Training duration (h)	3.417	17.405
Converge iterations	45000	15000

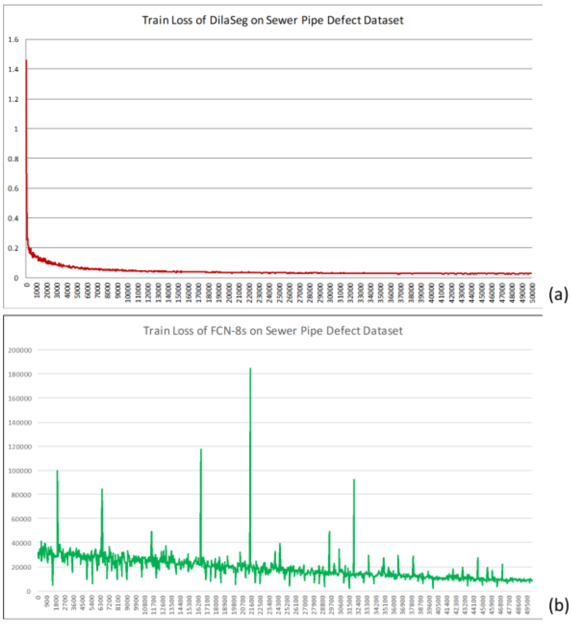


Figure 8. Training loss of DilaSeg and FCN-8s

5 Conclusion and Future Work

Computer vision techniques are attracting attention for automatic interpretation of sewer inspection results. Previous studies mainly focus on classifying and locating defects in images, which cannot provide information about the severity level of different defects. In addition, the requirement of the practical implementation of models on CCTV robots, e.g. the computational cost is rarely considered. This study aims to obtain the segmentation of three different types of sewer pipe defects i.e. crack, deposit and tree root, from CCTV inspection images to facilitate real-time severity assessment in the future.

To address the problem of large information loss during the down-sampling process of most previous deep learning models such as FCN, a new model called DilaSeg is developed in this study for semantic segmentation based on dilated convolution to increase feature map resolution. The proposed model is featured for performing several modules to obtain dense segmentation results, including normal convolution, dilated convolution, multi-scale dilated convolution as well as bilinear interpolation. Important features are extracted using the normal convolution and the feature maps are down-sampled due to consecutive convolution and max-pooling. Dilated convolution is implemented to prevent too much spatial information loss, which is also applied for objects with different scales through multi-scale dilated convolution. In the end, the feature maps are upsampled to original scale using bilinear interpolation. Experiments demonstrate that compared with FCN-8s, DilaSeg improved segmentation accuracy significantly in terms of all the evaluation indices. Especially, there is 18% of increase in mean PA and 22% of increase in mean IoU. Furthermore, the inference of the DilaSeg is faster than FCN-8s, which indicates the advantage of the proposed model for real-time application. Regardless of the improved accuracy, there are still some negative segmentations, possible reasons and potential solutions need to be validated in the future.

References

[1] EPA, Report to Congress on Impacts and Control of Combined Sewer Overflows and Sanitary Sewer Overflows, 2004.

[2] A. Arnab, S. Zheng, S. Jayasumana, B. Romera-Paredes, M. Larsson, A. Kirillov, B. Savchynskyy, C. Rother, F. Kahl, P.H.S. Torr, Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation: Combining Probabilistic Graphical Models with Deep Learning for Structured Prediction, IEEE Signal Process. Mag. 35 (2018) 37–52. doi:10.1109/MSP.2017.2762355.

[3] E. Shelhamer, J. Long, T. Darrell, E. Shelhamer, T. Darrell, J. Long, T. Darrell, E. Shelhamer, T. Darrell, Fully Convolutional Networks for Semantic Segmentation, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2015) 3431–3440. doi:10.1109/CVPR.2015.7298965.

[4] M.R. Halfawy, J. Hengmeechai, Automated defect detection in sewer closed circuit television images using histograms of oriented gradients and support vector machine, Autom. Constr. 38 (2014) 1–13. doi:10.1016/j.autcon.2013.10.012.

[5] S.S. Kumar, D.M. Abraham, M.R. Jahanshahi, T. Iseley, J. Starr, Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks, Autom. Constr. 91 (2018) 273–283. doi:10.1016/j.autcon.2018.03.028.

[6] J.C.P. Cheng, M. Wang, Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques, Autom. Constr. 95 (2018) 155–171. doi:10.1016/j.autcon.2018.08.006.

[7] S. Iyer, S.K. Sinha, Segmentation of pipe images for crack detection in buried sewers, Comput. Civ. Infrastruct. Eng. 21 (2006) 395–410. doi:10.1111/j.1467-8667.2006.00445.x.

[8] M. Thoma, A Survey of Semantic Segmentation, CoRR. (2016) 1–16. http://arxiv.org/abs/1602.06541.

[9] P. Krähenbühl, V. Koltun, Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials, Proc. 30th Int. Conf. Int. Conf. Mach. Learn. - Vol. 28. (2012) 513–521. doi:10.1.1.760.9549.

[10] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 2481–2495. doi:10.1109/TPAMI.2016.2644615.

[11] M. Holschneider, R. Kronland-Martinet, J. Morlet, P. Tchamitchian, A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform, in: J.-M. Combes, A. Grossmann, P. Tchamitchian (Eds.), Wavelets. Inverse Probl. Theor. Imaging, Springer Berlin Heidelberg, Berlin, Heidelberg, 1990: pp. 286–297.

[12] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, J. Garcia-Rodriguez, A Review on Deep Learning Techniques Applied to Semantic Segmentation, (2017) 1–23. http://arxiv.org/abs/1704.06857.

[13] A. Torralba, B.C. Russell, J. Yuen, LabelMe: Online Image Annotation and Applications, Proc. IEEE. 98 (2010) 1467–1484. doi:10.1109/JPROC.2010.2050290.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional Architecture for Fast Feature Embedding, in: Proc. 22Nd ACM Int. Conf. Multimed., ACM, New York, NY, USA, 2014: pp. 675–678. doi:10.1145/2647868.2654889.