# How Do Label Errors Affect Thin Crack Detection by DNNs

Liang Xu[1] Han Zou[1,2] Takayuki Okatani[1,2]
[1]Graduate School of Information Sciences, Tohoku University [2]RIKEN Center for AIP

TOHOKU UNIVERSITY

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

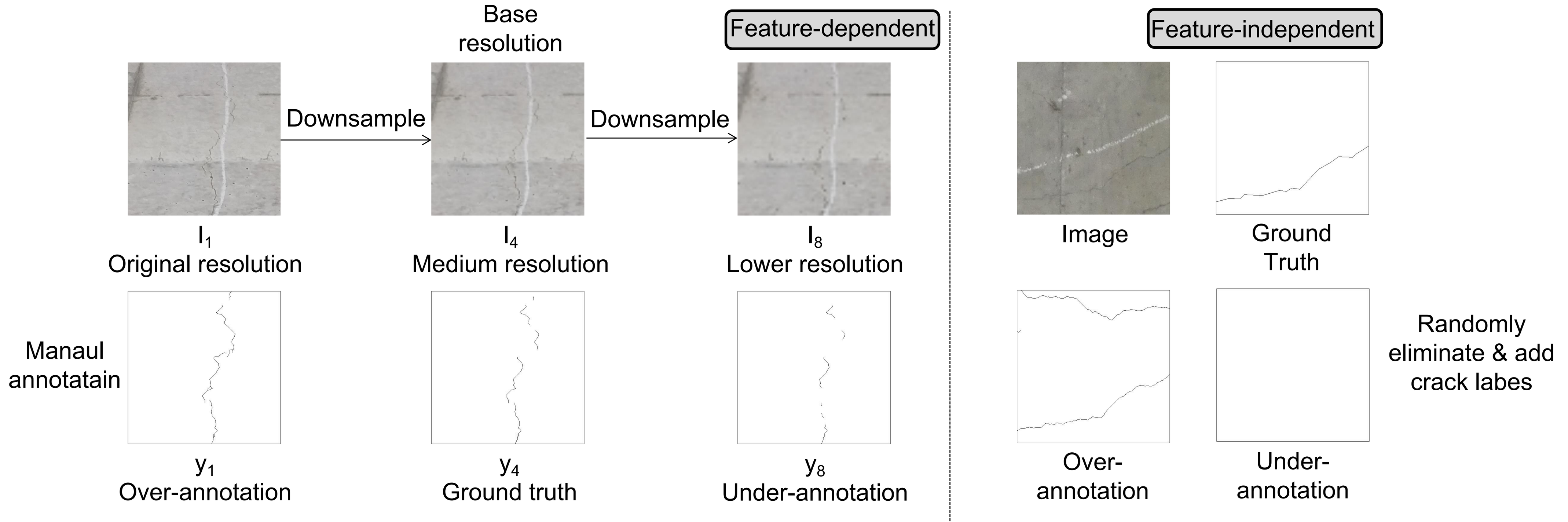
## 1. Problem

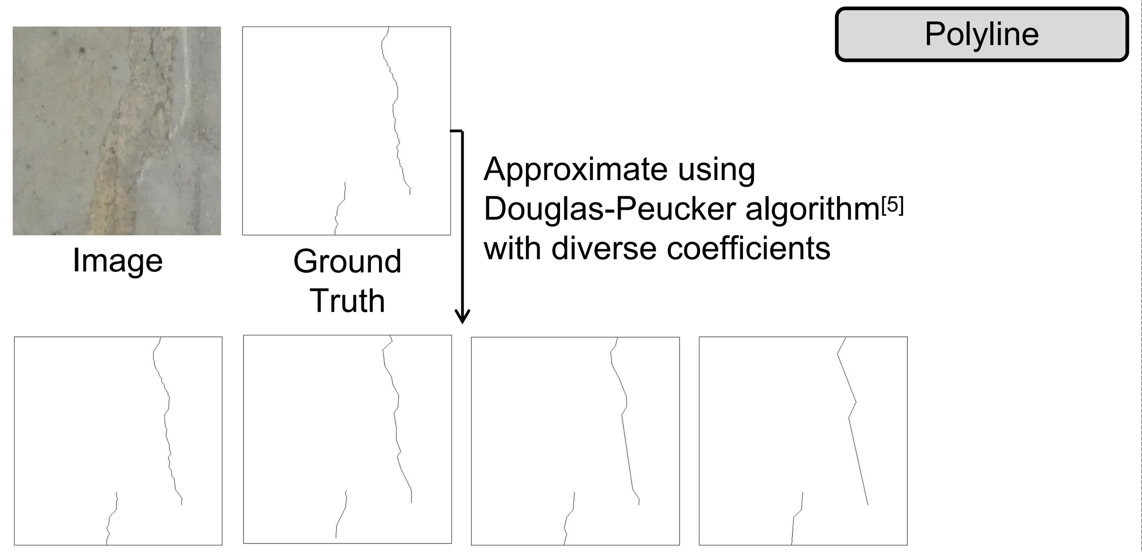
- Accurate annotation of cracks is crucial for supervised learning, but identifying thin cracks can be challenging;
- Simple yields effective annotation of crack curves can reduce annotation costs;
- little is known about how annotation errors in training data affect the accuracy of detectors trained on them.

## 2. Categorization:

- Feature-independent error emerge purely randomly annotation errors that emerge purely randomly;
- Feature-dependent annotation error usually results from some image structures;
- Polyline annotation error occurs when annotators do not provide the precise trace of a crack.

## 3. Synthesizing Noises:




### Experimental Configuration

512×512 pixel patches;
2,341 and 479 positive patches for training and testing, respectively;
Error levels veries from 10% to 50% in patch-wise fashion.
Mean squared error (MSE) loss
Implementation:Pytorch;
Augmentation: random 360-degree rotation and random flipping;
Training epoch: 80;
Learning rate: MultiStepLR;
Optimizer: Adam
Baseline models:UNet[2], HRNet-W18-C[3], DeepCrack[4], and Crackformer[5]

## 4 Result

After training the UNet model on the training data with synthetic errors, we evaluate its performance on error-free images from the test set.

- Effects of Feature-independent Errors.

| Error level | Clean 0 | Under-annotation 10 | 20 | 30 | 40 |
|---|---|---|---|---|---|
| HRNet | 0.950 | 0.937 | 0.921 | 0.894 | 0.859 |
| CrackFormer | 0.929 | 0.914 | 0.894 | 0.861 | 0.817 |
| DeepCrack | 0.903 | 0.885 | 0.859 | 0.821 | 0.762 |
| UNet | 0.882 | 0.861 | 0.830 | 0.786 | 0.710 |
| **Error level** | **Clean 0** | **Over-annotation 10** | **20** | **30** | **40** |
| HRNet | 0.950 | 0.943 | 0.932 | 0.915 | 0.884 |
| CrackFormer | 0.929 | 0.921 | 0.907 | 0.885 | 0.846 |
| DeepCrack | 0.903 | 0.891 | 0.872 | 0.841 | 0.789 |
| UNet | 0.882 | 0.867 | 0.843 | 0.804 | 0.742 |

Table 1. F1-scores on error-free test images of different DNN models trained with training data having different levels of annotation errors.

- Detection performance deteriorates with increasing levels of errors, but the impact is modest and tolerable with low error levels.
- Comparing under- and over-annotation errors, the impact is smaller for over-annotation.
- The impact of 30% error for over-annotation is mostly the same as that of 20% error for under-annotation

- Effects of Feature-dependent Errors.

| Error level | Clean 0 | Under-annotation 10 | 20 | 30 | 40 |
|---|---|---|---|---|---|
| HRNet | 0.950 | 0.935 | 0.889 | 0.848 | 0.785 |
| CrackFormer | 0.929 | 0.911 | 0.884 | 0.832 | 0.729 |
| DeepCrack | 0.903 | 0.879 | 0.843 | 0.764 | 0.637 |
| UNet | 0.882 | 0.854 | 0.811 | 0.713 | 0.563 |
| **Error level** | **Clean 0** | **Over-annotation 10** | **20** | **30** | **40** |
| HRNet | 0.950 | 0.948 | 0.943 | 0.931 | 0.914 |
| CrackFormer | 0.929 | 0.924 | 0.916 | 0.901 | 0.882 |
| DeepCrack | 0.903 | 0.897 | 0.886 | 0.864 | 0.837 |
| UNet | 0.882 | 0.875 | 0.862 | 0.836 | 0.803 |

Table 2. F1-scores achieved by diverse DNN models trained with various types of feature-dependent noisy labels under different error levels.

- When compared to error-free cases, 40% under-annotation errors lead to a considerable reduction in F1-scores for HRNet, CrackFormer, DeepCrack, and UNet, with a reduction of 17.3%, 21.5%, 29.5%, and 36.2%, respectively.
- 40% over-annotation errors only result in a minor decrease of 3.9%, 5.1%, 7.3%, and 9.0%, respectively.

- Effects of Mix of Under- and Over-annotation.

| Combination($\widetilde{\mathcal{R}}_u/\widetilde{\mathcal{R}}_o$) | 30/0 | 20/10 | 15/15 | 10/20 | 0/30 |
|---|---|---|---|---|---|
| Independent | 0.786 | 0.793 | 0.797 | 0.800 | 0.804 |
| Dependent | 0.713 | 0.786 | 0.842 | 0.854 | 0.836 |

Table 3. F1-scores of the UNet model trained with data having dif_x0002_ferent ratios of under- and over-annotation errors whose total error level is 30%.

| Combination(Ru/Ro) | 40/0 | 30/10 | 20/20 | 10/30 | 0/40 |
|---|---|---|---|---|---|
| Independent | 0.710 | 0.721 | 0.728 | 0.734 | 0.742 |
| Dependent | 0.563 | 0.759 | 0.810 | 0.829 | 0.803 |

Table 4. F1-scores of the UNet model trained with data having dif_x0002_ferent ratios of under- and over-annotation errors whose total error level is 40%.

- In the case of **feature-independent** errors, the ratio of under- and over-annotation has no significant impact on detection accuracy, while the total error level is the main factor.
- In the case of feature-dependent errors, the ratio of under- and over-annotation has a significant impact.
- A combination of under- and over-annotation errors is likely to improve accuracy.
- it is crucial for annotators to ensure that under-annotation errors occur less frequently than over-annotation errors.

- Effects of Polyline Annotation

| Comp. rate(%) | HRNet | CrackFormer | DeepCrack | UNet |
|---|---|---|---|---|
| 100 | 0.950 | 0.929 | 0.903 | 0.882 |
| 75 | 0.944 | 0.925 | 0.898 | 0.878 |
| 50 | 0.938 | 0.919 | 0.892 | 0.875 |
| 25 | 0.931 | 0.912 | 0.884 | 0.868 |
| 10 | 0.920 | 0.903 | 0.877 | 0.860 |
| 5 | 0.908 | 0.895 | 0.868 | 0.849 |
| 3 | 0.897 | 0.883 | 0.854 | 0.831 |
| 1 | 0.883 | 0.872 | 0.826 | 0.792 |

Table 5. F1-scores achieved by the four methods trained on polyline annotation with different compression rates.

- the accuracy of the models decreases proportionally to the compression rate; their ranking remains unchanged regardless of the compression rate.
- It should be noted that even at a compression rate of 25%, all the models show only a modest decrease in accuracy.

## 5.Examples

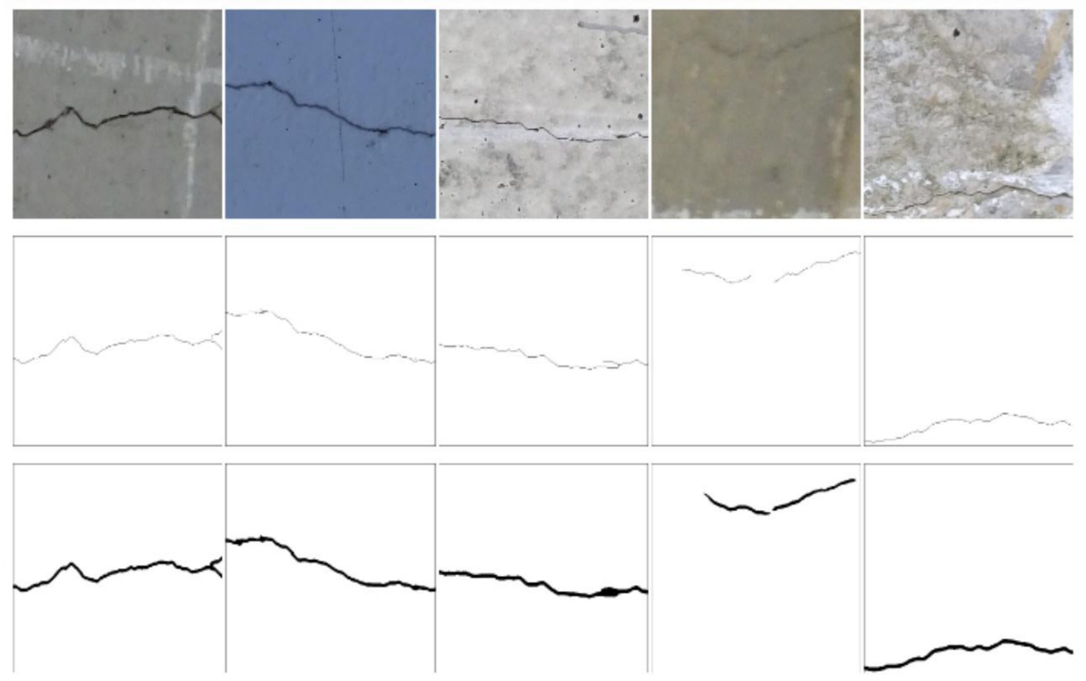


Figure 1. Examples of detection result. The first row is the image, the second row is the label and the third row is the output from our model.

## 6.Summary/Conclusion

- We defined assessed three types of errors that affect crack detection accuracy.
- We experimentally found that under-annotation has a more significant negative impact than over-annotation.
- We also found combination of under- and over-annotation errors tend to improve accuracy.
- We finally found that polyline annotation is an effective way to reduce annotation costs while maintaining model detection accuracy.

### References

[1] John Hershberger and Jack Snoeyink. Speeding up the douglas-peucker line-simplification algorithm. In Proc. 5th Intl. Symp. on Spatial Data Handling, pages 134–143, 1992.

[2] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[3] Changqian Yu, Bin Xiao, Changxin Gao, Lu Yuan, Lei Zhang, Nong Sang, and Jingdong Wang. Lite-hrnet: A lightweight high-resolution network. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10440–10450, 2021.

[4] Yahui Liu, Jian Yao, X. Lu, Renping Xie, and L. Li. Deepcrack: A deep hierarchical feature learning architecture for crack segmentation. Neurocomputing, 2019.

[5] Z. Lu, Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao. Learning from weak and noisy labels for semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017.