Active Learning for Imbalanced Civil Infrastructure Data

Thomas Frick[1,2], Diego Antognini[1], Mattia Rigotti[1], Ioana Giurgiu[1], Benjamin Grewe[2], and Cristiano Malossi[1]

[1] IBM Research, Zurich, Switzerland
{fri,mrg,igi,acm}@zurich.ibm.com, diego.antognini@ibm.com,
[2]
Institute of Neuroinformatics, UZH and ETH Zurich, Switzerland
bgrewe@ethz.ch

Abstract. Aging civil infrastructures are closely monitored by engineers for damage and critical defects. As the manual inspection of such large structures is costly and time-consuming, we are working towards fully automating the visual inspections to support the prioritization of maintenance activities. To that end we combine recent advances in drone technology and deep learning. Unfortunately, annotation costs are incredibly high as our proprietary civil engineering dataset must be annotated by highly trained engineers. Active learning is, therefore, a valuable tool to optimize the trade-off between model performance and annotation costs. Our use-case differs from the classical active learning setting as our dataset suffers from heavy class imbalance and consists of a much larger already labeled data pool than other active learning research. We present a novel method capable of operating in this challenging setting by replacing the traditional active learning acquisition function with an auxiliary binary discriminator. We experimentally show that our novel method outperforms the best-performing traditional active learning method (BALD) by 5% and 38% accuracy on CIFAR-10 and our proprietary dataset respectively.

Keywords: Class Imbalance, Active Learning, Deep Learning, Neural Networks

1 Introduction

Civil infrastructures are constantly being monitored for critical defects, as failure to recognize deficiencies can end in disaster. Unfortunately, detailed inspections are time-consuming, costly, and sometimes dangerous for inspection personnel. We are working towards automating all aspects of visual inspections of civil infrastructures by recent advances in drone technology [18, 6] that enable fast and remote visual inspection of inaccessible structures such as wind turbines, water dams, or bridges and simultaneous advances in instance segmentation using deep learning models [28, 8, 36] that facilitate accurate detection of surface defects on high resolution images. We are combining these novel technologies by utilizing

drones equipped with high-resolution cameras to capture image material of the structures and deep neural networks to detect defects. Our goal is to decrease the duration of inspections, improve condition assessment frequency, and ultimately minimize harm to human health.

It is relatively simple to collect large amounts of data of civil infrastructures for our proprietary dataset. However, similar to many other projects applying deep learning, annotating the collected samples is incredibly costly and time-intensive as highly-trained engineers are required for the labeling process. To minimize costs, a popular approach is to annotate a small portion of the dataset first and then use Active Learning (AL) to select informative samples that should be labeled next. The goal is to optimize a trade-off between additional annotation cost (number of annotated samples) and an increase in model performance. AL has been successfully applied in medical imaging [33, 24], astronomy [29], and surface defect detection [9]. For its application in industry, we observe two differences to the traditional AL setting:

First, previous research[11, 3, 20] has focused on starting the AL process with little to no data (100 samples or cold start). In contrast, industry projects often start the AL process with a larger pool of labeled training data: Despite the high annotation costs, a random set of initial data is labeled for a proof of concept. Only then data collection and labeling efforts are scaled up. In this phase, the efficient selection of samples is invaluable. Unfortunately, traditional AL strategies barely outperform random selection given a large initial dataset, as shown in our experiments.

Secondly, most academic datasets are class balanced, with each class having the same number of samples (e.g., CIFAR-10 [21] 5,000). In contrast, real-world industry datasets usually suffer from a long-tailed class distribution [39, 25]. Moreover, the minority classes are often the most important ones. This is usually the case for civil structures: dangerous/critical defects rarely appear on drone scans of the structure as they are well maintained. Previous work on AL for imbalanced data [4, 34, 22] has shown that classical AL methods fail to select samples of the minority classes for heavily imbalanced datasets and therefore fail to improve model performance for the minority classes. These works focus on developing sample selection strategies that choose samples according to an uncertainty-based metric or a diversity-based approach.

We present a novel method that effectively and efficiently selects minority samples from a pool of unlabeled data for large datasets suffering from heavy class imbalance. Contrary to other AL methods that try to find informative samples from all classes, we limit ourselves to selecting samples for a single minority class, investing the total labeling budget for samples that improve model performance for only that minority class. To this end, our method replaces the AL acquisition function with a binary discriminator explicitly trained in a one-vs-all fashion (minority vs. majority classes) to distinguish between unlabeled minority and majority samples. In each cycle, the discriminator selects samples to be labeled next according to the highest prediction scores. We experimentally confirm that classical active learning methods fail to significantly improve model

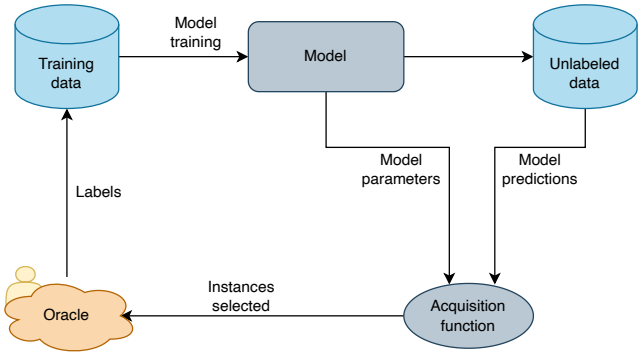


Fig. 1: Active Learning cycle: a model is trained on a pool of labeled data. Given the model predictions and parameters, an acquisition function selects informative samples from an unlabeled pool of data which are sent to the oracle for labeling.

performance for CIFAR-10 [21] and our proprietary civil engineering dataset 3. Applying our method to our proprietary civil infrastructure dataset, we show a minority class recall improvement of 32% and an overall accuracy gain of 14% compared to the best-performing traditional AL method (BALD [17]).

2 Related Work

Active learning is a well studied problem [31, 13, 4] and especially advantageous for applications with high annotation costs because highly specialized annotators are needed. Active learning works by iterative selecting informative samples according to a query strategy from an unlabeled pool of data. The chosen samples are passed to the oracle (usually a human annotator) to be labeled. The goal is to improve model performance as much as possible while labeling as few samples as necessary. Settles et al. [31] identifies two different AL scenarios: In sample-based selective sampling, a stream of samples arrives one after another. Therefore, the algorithm decides whether to label or discard each sample without information about samples arriving in the future. In contrast, in pool-based active learning, the algorithm has access to the entire pool of unlabeled samples and needs to select samples to be labeled next. Our work is set in the second scenario.

Active Learning query strategies can be divided into geometric-based and uncertainty-based methods. Geometric-based approaches make use of the underlying feature space to select informative samples. Feature embeddings are necessary to have a meaningful feature space for an application to high-dimensional input data such as images. For our application, we focus on uncertainty-based methods as we assume that there is no pre-trained model available that can be used as an embedding. There is a wide variety of work on uncertainty estimation

of deep neural networks [12]. Two popular methods that estimate uncertainty by generating multiple predictions on the same input are ensembling multiple models [23] or using Dropout as a bayesian approximation [11]. As the first option is highly compute-intensive, we make use of the second one for this work.

Uncertainty-based acquisition functions judge a samples informativeness based on model uncertainty. The Entropy acquisition function [32] chooses samples that maximise the predictive entropy, BALD [17] selects data points that are expected to maximise the information gained about the model parameters, Variation Ratios [10] measures lack of confidence. BatchBALD [20] improves on BALD by selecting sets of samples that are jointly informative instead of choosing data points that are informative individually. As BatchBALD is impractically compute-intensive for large batches and BALD has been shown to outperform other older methods, we focus only on Entropy and BALD for our experiments.

Imbalanced datasets have been widely studied (see Kaur et al. [19] for an extensive overview). Prior work frequently utilizes resampling techniques (e.g., oversampling and undersampling, or a combination of both) to balance the training data. While Hernandez et al. [16] show that simple resampling techniques can significantly improve model performance, Mohammed et al. [27] conclude that undersampling may discard informative majority class samples and therefore decrease majority class performance. There are synthetic sampling methods (SMOTE [7], ADSYN [14]) that generate new samples by interpolating in the underlying feature space. However, the underlying input feature space is too complex for image data to apply these methods successfully. Weighting samples according to inverse class frequency has also been successfully applied to prioritize underrepresented classes in training. For our work, we exclusively use oversampling as combining duplication of samples with image augmentation techniques results in a much more diverse set of minority samples than weighting the classes. One disadvantage of this method is the resulting much larger training pool and the longer training times compared to the class-weighted approach.

3 Background

As mentioned above, we work towards automating visual inspections of civil infrastructures. We have developed an automated inspection pipeline consisting of 4 consecutive stages: First, drones capture grids of high-resolution images, which are later stitched into full image scenes. Next, state-of-the-art deep learning methods analyze the image scenes and highlight defects such as cracks or rust. Finally, the size of detected defects is measured with a precision of 0.1mm. The pipeline's output supports civil engineers in prioritizing further in-person inspections and maintenance activities.

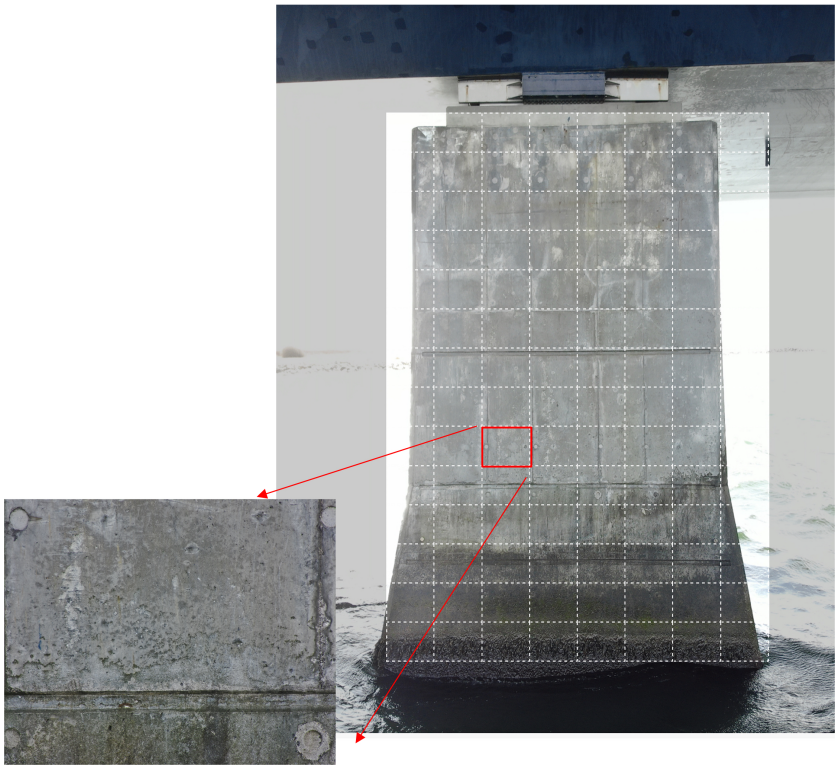


Fig. 2: DJI Matrix 300 Hi-Res Grid: example of a 7-grid of 98 high-resolution images. Each individual image is 5184×3888 pixel in size resulting in a overall data size for a single pillar face of almost 1GB.

3.1 Image Data

We have collected a civil infrastructure dataset consisting of high-resolution images of concrete bridge pillars. The data was collected in 2021 by certified drone pilots using DJI Matrix 300 [1] drones equipped with a Zenmuse H20 [2] lens attachment. Scenes were captured using DJI's Hi-Res Grid Photo mode as shown in Fig. 2. Drone pilots manually mark the area of the civil structure under inspection on an overview image. While hovering in place, the drone captures a grid of overlapping images by gimbaling the zoom lens. In total, 22 bridge pillars were scanned from each side, resulting in a dataset of over 22'000 raw images (5184×3888 pixels).

3.2 Instance Segmentation Annotations

In light of our focus on defect detection, we have invested considerable effort in creating a high-quality instance segmentation dataset where each defect is categorized and localized using mask annotations. In collaboration with our civil

engineering experts, we developed extensive annotation guidelines focusing on the following six defect categories: rust, spalling, cracks, cracks with precipitation, net-cracks, and algae (see Fig. 3 for examples). Over six months, a team of annotators labeled 2500 images resulting in around 14'000 defect annotations. Given the well-maintained condition of the inspected structure, critical defects such as rust or cracks with precipitation appear infrequently. Consequently, our dataset suffers from a long-tailed class distribution with a maximum imbalance ratio of 1 to 130 (cracks with precipitation vs. cracks).

4 Method

This section first describes our novel method which replace the acquisition function with a binary discriminatory trained on the labeled pool to select minority class samples from the unlabeled pool. Secondly, it describes the dataset conversion algorithm which we use to convert our proprietary instance segmentation dataset to a classification dataset for the experiments.

4.1 Active Learning for heavily imbalanced data

The goal of the classical AL setting is to improve the model performance by labeling additional informative samples. A model is trained on the initial pool of labeled data. Given the model's predictions and parameters, an acquisition function determines which samples from the unlabeled pool are sent to the oracle for labeling. The newly annotated samples are then added to the labeled pool, after which the cycle begins again by retraining the model. Traditional AL methods [11, 17, 3, 20] strive to select samples for which the model performance for all classes improves most.

In contrast, we focus on improving model performance for a single preselected minority class. Given the high class imbalance of the initial AL dataset, we hypothesize that we do not need to find the most diverse or informative set of samples but that selecting and labeling any samples of the minority class will improve model performance. Therefore, our method replaces the traditional uncertainty-based AL acquisition function with an auxiliary binary classifier acting as a discriminator between the minority and majority classes. The discriminator is trained on the labeled AL pool in each cycle. Its predictions on the unlabeled pool are used to select samples for labeling. Selection is based on prediction scores, choosing the top-K samples.

For the training of the binary discriminator, binary labels are computed from the original multi-class training set in a one-vs-all fashion. As a result, the binary classification dataset is even more imbalanced than the original multi-class dataset. We improve discriminator performance by combating the bias of the class imbalance by applying the following modifications to the training procedure:

- oversample the positive class (the original minority class) until we reach balance with respect to the negative class (all majority classes);

– apply standard image augmentation techniques (flip, shift, scale, rotate, brightness, and contrast - see Albumentations [5]). As we apply the augmentations to the large number of minority samples generated from the oversampling in the first step, we end up with a highly diverse set of minority class samples;
– apply batch augmentations (MixUp [38], CutMix [37] - see timm [35] library) to further diversify the samples in each batch stabilizing the training procedure of the binary discriminator.

4.2 Instance segmentation to classification dataset conversion

As mentioned in section 3, we are working towards automated detection of surface defects for civil infrastructures. We train instance segmentation models on the original dataset with instance mask annotations to detect the defects. Our annotations must be labeled with extraordinary precision due to the defects being of such small size (e.g., cracks in the order of millimeters). This additional time effort adds to the already high prize per annotation. Therefore, we have been working on a technique to use weakly supervised learning with class-level supervision to generate class activation maps (CAMs [30]) from which we extract fine segmentation masks. Unfortunately, there are still high annotation costs associated with class-level labels. We are trying to decrease these costs with active learning in this work. Consequently, we use class-level labels for our experiments which we extract from the original instance segmentation dataset using a patch-based dataset conversion algorithm.

The conversion algorithm works by extracting fixed-size patches from the original images. Intuitively, a patch can be assigned to a category if it depicts a piece of the original class; this can be either shape or texture. While the texture is informative enough for some objects to attribute the correct class, others are only correctly classifiable with information about their shape. As extracted patches only offer a small window into the original image, an object's shape information can be easily lost if a too small patch size is chosen.

The algorithm processes one image at a time: multiple patches are sampled for each instance and assigned the corresponding category. Additionally, patches are extracted from regions of the original image that do not contain any instances/defects and are assigned to an additional, newly introduced class "Background". We randomly sample instance patches such that the center point of the patch lies within the instance annotation and background patches such that the center point lies anywhere within the original image's borders. The algorithm rejects a sampled patch if it violates one of the following criteria: 1) it overlaps with an already chosen valid patch, 2) it intersects with an instance annotation belonging to a different category, 3) the patch breaches the boundaries of the original image. The algorithm extracts patches until the total number of required patches is reached or until the algorithm exceeds a total number of sampling attempts. Figure 3 shows the sampled patches for an example image as well as one example patch per category.

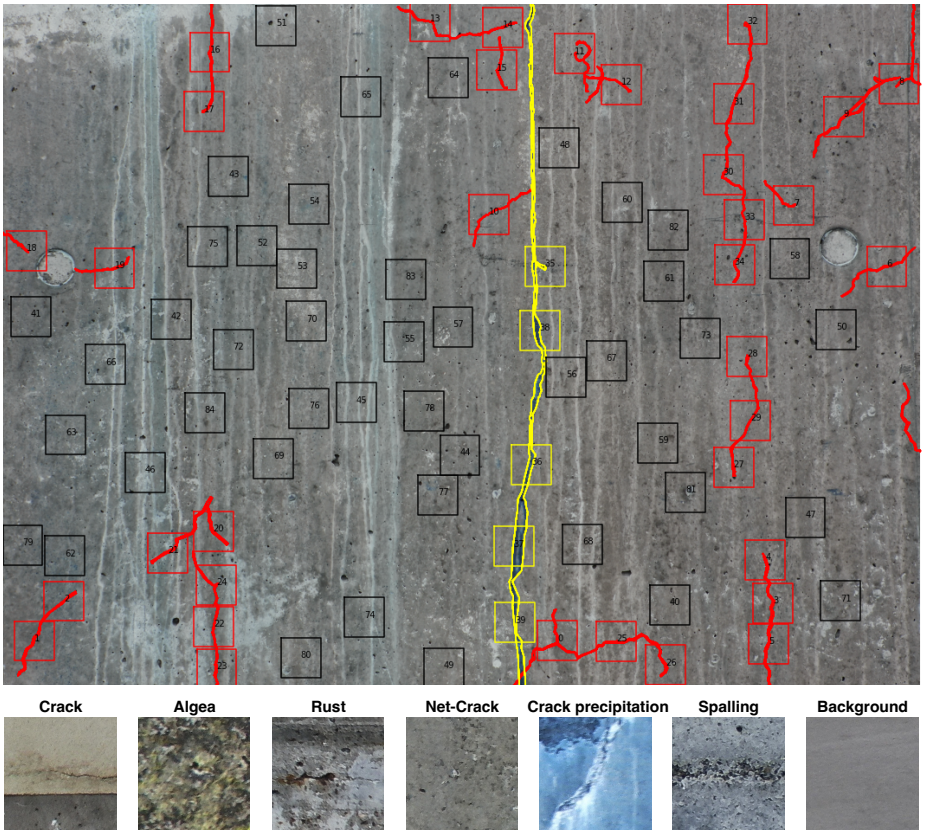
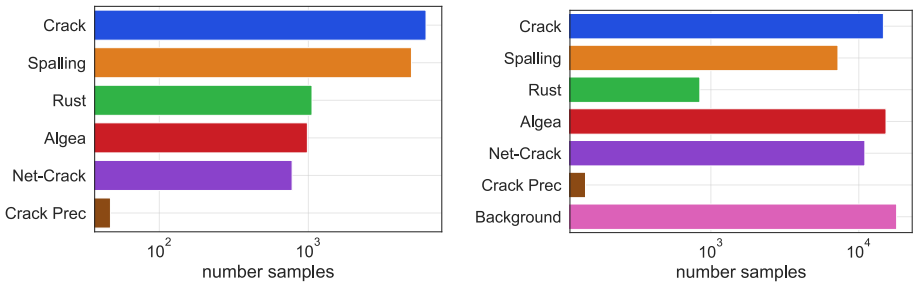


Fig. 3: Patch-based classification dataset conversion algorithm: we show instance segmentation polygons of the original image colored by class (red crack, yellow crack with precipitation, black background) and sampled patch boarders with the same color scheme. Below, we show one example patch per category of the original dataset plus one example patch for the additional background class.

5 Evaluation

In this section we first describe how we convert the image segmentation to a classification dataset with our patch extraction method to prepare our proprietary dataset. Then, we experimentally show how traditional active learning methods fail to improve minority class performance as the initial training data size increases. Finally, we run experiments showcasing model performance for traditional active learning methods, as well as our proposed method on CIFAR-10 and on the civil engineering classification dataset.



(a) Instance segmentation dataset (b) Extracted classification dataset

Fig. 4: Class distribution of our civil infrastructure datasets: (a) shows statistics for our original instance segmentation dataset while (b) describes statistics for the classification dataset extracted with the path-based algorithm from the instance segmentation dataset (a). The additional background class in Figure (a) is a result of the background patch sampling of the algorithm.

5.1 Civil infrastructure classification dataset

We apply our instance segmentation to classification dataset conversion algorithm (see section 4.2), extracting 160×160 pixel patches. Per image, the algorithm attempts to sample 100 class-patches and 10 background-patches with a maximum of 100 sampling attempts. The resulting classification dataset consists of a total of 67'162 samples. Fig. 4 shows the class distribution for the original instance segmentation dataset and for the extracted classification dataset. Both suffer from heavy class imbalance.

We need a large enough dataset for our experiments to simulate a real-world active learning scenario with a small labeled and a large unlabeled pool. Accordingly, each category included in the final dataset has to contain enough samples such that it is possible to introduce artificial imbalance for each class during our experiments. Therefore, we remove all classes with less than 5'000 samples (crack with precipitation, rust). Additionally, we remove the net-crack class due to the visual similarity of its patches with the crack class. The final dataset consists of four classes: background, algae, crack, and spalling. It is split in a stratified fashion into 70% training set and 30% test set.

5.2 Experiment setup

Active learning datasets We aim to simulate a real-world active learning scenario with a small pool of labeled data and a large pool of unlabeled data, as well as an oracle that can be queried for labels. Artificial class imbalance is introduced into the labeled and the unlabeled pool to simulate the imbalanced dataset setting dependent on the experiment. For the experiments, the original training set is consequently randomly split into a small labeled set, a large unlabeled set, and an unused set. As we still have access to the labels of the samples

of the unlabeled pool, we can simulate human annotation when the active learning algorithm queries labels. Furthermore, once samples have been moved from the unlabeled pool to the labeled pool (simulated labeling process), we can simulate a much larger unlabeled pool by moving the same number of samples per class from the unused pool to the unlabeled pool, restoring the original unlabeled data pool size and class balance. Finally, the test set is used as-is to evaluate the models after each time the AL algorithm queries new samples.

Model training We use the full ResNet18 [15, 35] as a model backbone for our civil infrastructure dataset, but only use the first ResNet block for CIFAR-10 [21]. All models are trained for 50 epochs to convergence with the AdamW [26] optimizer with a learning rate of 0.0005 with the sample and batch augmentations highlighted above. For a fair comparison between the traditional AL methods and ours, we train all models with the improved training procedure of oversampling, image and batch augmentations.

Active learning process We apply our method and traditional AL algorithms to each dataset under investigation. AL algorithms query labels five times for 200 samples per cycle. Each experiment is repeated three times selecting a class from the original dataset as the minority class for the artificially introduced imbalance procedure. All experiments use different randomly initialised model weights and a different random split of the data (labeled and unlabeled pool with artificial imbalance).

5.3 Results

Influence of initial training pool size As a first experiment we evaluate all methods on CIFAR-10 [21] with increasing number of samples in the initial labeled pool. We create artificially imbalanced training datasets with a increasing number of samples for the majority class from 500 to 1500 while keeping the number of minority class samples constant at 50. To minimize the impact of the unlabeled set, it is kept constant with 120 minority samples and 3000 majority samples. Finally, we run the AL cycle for five rounds, selecting and labeling 200 samples per cycle. We measure the change in performance from the initial trained model to the model at the end of the AL procedure. As can be seen in Fig. 5, the delta in minority class performance drops off significantly for traditional active learning methods as the initial training dataset size increases. In contrast, our method retains more of the original performance gains as it is less sensitive to the initial data pool size. Specifically, the recall delta remains above zero as initial data size decreases while the classical AL method fall below zero, signaling a decrease in performance from the initial training set to the set with the additional 1000 labeled samples.

CIFAR-10 Next, we evaluate absolute model performance on CIFAR-10 [21], comparing traditional AL methods (BALD [17], Entropy [32]) with our novel

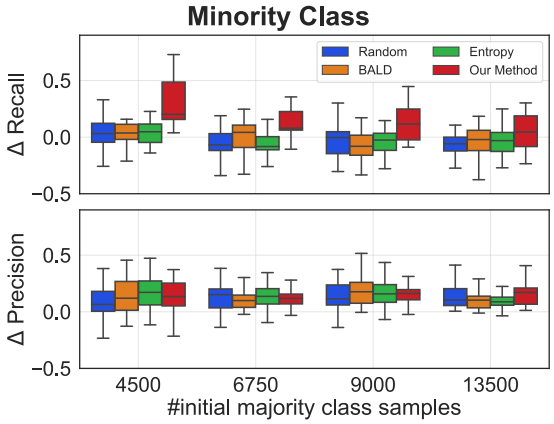
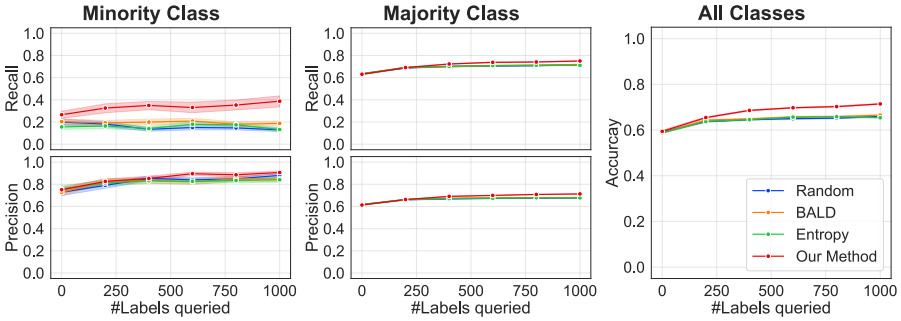
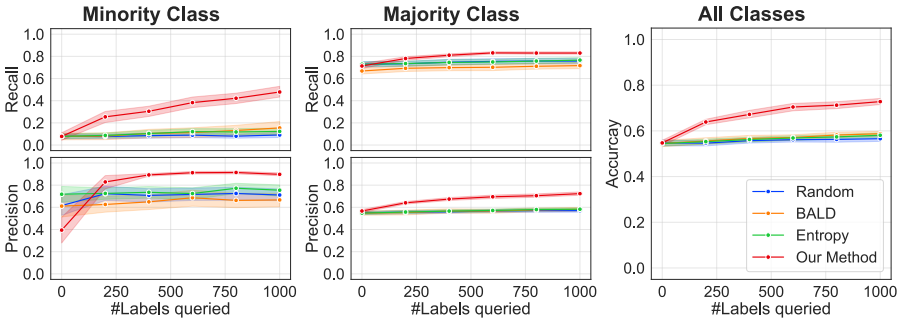


Fig. 5: Influence of initial AL training pool size: we report relative minority class performance gains for four dataset sizes with an increasing number of majority samples in the initial AL training pool. Performance differences are evaluated for each method, comparing model performance before the AL procedure with model performance after the procedure queried labels for 1000 samples. While performance gains of traditional AL methods decrease as the training dataset grows, our method retains more of the original performance increase.

method on model performance. Additionally, we also include a random acquisition function as a baseline. We create an initial AL dataset that is randomly split into unlabeled and labeled pool for each experiment with an artificial class imbalance of 50 minority samples to 1000 majority samples per class. The unlabeled pool consists of 300 minority samples and 3000 majority samples per class. Results in Fig. 6a show that traditional AL methods fail to significantly improve precision and recall for the minority class. Additionally, little performance improvement can be seen for the majority classes. We explain this with our experiments' much larger initial training pool compared to other publications. Our initial training set consists of few minority samples but a considerable amount of majority samples. The 200 additionally labeled samples per cycle do not yield much additional information compared to the existing larger training pool. Therefore, neither majority nor minority class performance improves. In contrast, our method focuses only on the minority class, for which only a few samples are in the initial training pool. Therefore, even a few minority samples yield enough information to improve minority class performance considerably. Our method shows a clear performance improvement compared to traditional AL methods. Compared to the random baseline, our method improves on average by 25% recall of the minority class, while BALD only improves by 5% and the entropy method only improves by 0.3%. As a result, the overall accuracy of the model also increases significantly: 5% average improvement over the random



(a) CIFAR-10



(b) Civil engineering dataset

Fig. 6: Absolute model performance throughout the AL process: for each cycle, after labeling 200 additional samples, we report precision and recall for the minority class, macro average precision and recall for the majority classes, and overall accuracy for the CIFAR-10 dataset and our proprietary civil infrastructure dataset. Error bands show the standard error of the mean (SEM).

baseline for our method, compared to only 0.6% for BALD and a 0.5% decrease for the entropy method.

Civil engineering dataset We evaluate absolute model performance on our proprietary civil engineering dataset (see section 5.1). We run experiments on an initial AL dataset with 50 minority and 1000 majority samples per class in the training set. The unlabeled pool consists of 300 minority and 2500 majority samples per class. As with CIFAR-10, the results in Fig. 6b show that traditional AL methods fail to improve precision and recall for the minority class. Meanwhile, our method shows clear improvement in minority class recall and precision as well as overall accuracy. Compared to the results on CIFAR-10, we see a larger performance improvement: compared to the random baseline our method improves by 38% minority recall, 18% minority precision, and 16% over-

all accuracy. BALD only improves by 6% minority recall and 2% overall accuracy while minority precision decreases by 4%. The entropy method improves only by 3% minority recall, 4% minority precision and 1% overall accuracy. We explain the higher overall accuracy gains compared to CIFAR-10 with the smaller number of classes in our dataset (4 compared to 10 in CIFAR-10). Given the overall accuracy is an average of class-specific performance, minority class improvement has a much larger impact on the majority classes.

6 Conclusion

We have presented a novel active learning method that replaces the traditional acquisition function with an auxiliary binary discriminator allowing the selection of minority samples even for initially large and imbalanced datasets. We have experimentally shown that our method outperforms classical AL algorithms on artificially imbalanced versions of CIFAR-10 and our proprietary civil engineering dataset when evaluated on minority class recall, precision, and overall classification accuracy. Consequently, our method facilitates the successful discovery and labeling of rare defects in the yet unlabeled pool of samples for our proprietary civil engineering dataset. Trained on the additional labeled data, our visual inspection defect detection models improve at supporting civil engineers' maintenance prioritization decisions for rare but critical defects.

Due to the large number of models trained per experiment, we are limited to dataset consisting of small images with many classes or large images with few classes. Additionally, our choices of datasets for experimentation were limited as we required many samples per class to simulate a large unlabeled pool. This excluded many popular large datasets as they often consist of many classes with a moderate amount of samples per class.

While we focus on classification datasets in this work, future research could extend our method to an application for the full instance segmentation dataset by creating sliding window patches and aggregating statistics over the full original image. Additionally, it would be interesting to develop our method further to work with multiple minority classes at a time. Finally, research should compare our method with feature space diversity-based AL methods when a pre-trained model (transfer learning or self-supervised learning) is available.

Acknowledgement This work would not have been possible without Finn Bormlund and Svend Gjerding from Sund&Bælt. We would like to thank them for their collaboration, specifically for the collection of image data, for their expert annotations, and their tireless help with the annotation guidelines for the civil engineering dataset.

References

1. Matrice 300 RTK – Built Tough. Works Smart., https://www.dji.com/ch/photo
2. Zenmuse H20 Series – Unleash the Power of One, https://www.dji.com/ch/photo
3. Bayesian Active Learning (BaaL) (Jul 2022), https://github.com/baal-org/baal, original-date: 2019-09-30T20:16:26Z
4. Aggarwal, U., Popescu, A., Hudelot, C.: Active Learning for Imbalanced Datasets. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1417–1426. IEEE, Snowmass Village, CO, USA (Mar 2020). https://doi.org/10.1109/WACV45572.2020.9093475, https://ieeexplore.ieee.org/document/9093475/
5. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: Fast and Flexible Image Augmentations. Information **11**(2), 125 (Feb 2020). https://doi.org/10.3390/info11020125, https://www.mdpi.com/2078-2489/11/2/125, number: 2 Publisher: Multidisciplinary Digital Publishing Institute
6. Chan, K.W., Nirmal, U., Cheaw, W.G.: Progress on drone technology and their applications: A comprehensive review. AIP Conference Proceedings **2030**(1), 020308 (Nov 2018). https://doi.org/10.1063/1.5066949, https://aip.scitation.org/doi/abs/10.1063/1.5066949, publisher: American Institute of Physics
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research **16**, 321–357 (Jun 2002). https://doi.org/10.1613/jair.953, https://www.jair.org/index.php/jair/article/view/10302
8. Davtalab, O., Kazemian, A., Yuan, X., Khoshnevis, B.: Automated inspection in robotic additive manufacturing using deep learning for layer deformation detection. Journal of Intelligent Manufacturing **33**(3), 771–784 (Mar 2022). https://doi.org/10.1007/s10845-020-01684-w, https://doi.org/10.1007/s10845-020-01684-w
9. Feng, C., Liu, M.Y., Kao, C.C., Lee, T.Y.: Deep Active Learning for Civil Infrastructure Defect Detection and Classification. In: Computing in Civil Engineering 2017. pp. 298–306. American Society of Civil Engineers, Seattle, Washington (Jun 2017). https://doi.org/10.1061/9780784480823.036, http://ascelibrary.org/doi/10.1061/9780784480823.036
10. Freeman, L.C.: Elementary Applied Statistics: For Students in Behavioral Science. Wiley (1965), google-Books-ID: r4VRAAAAMAAJ
11. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (Oct 2016), http://arxiv.org/abs/1506.02142, arXiv:1506.02142 [cs, stat]
12. Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X.: A Survey of Uncertainty in Deep Neural Networks (Jan 2022), http://arxiv.org/abs/2107.03342, arXiv:2107.03342 [cs, stat]
13. Hanneke, S.: Theory of Disagreement-Based Active Learning. Foundations and Trends® in Machine Learning **7**(2-3), 131–309 (Jun 2014). https://doi.org/10.1561/2200000037, https://www.nowpublishers.com/article/Details/MAL-037, publisher: Now Publishers, Inc.

14. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). pp. 1322–1328 (Jun 2008). https://doi.org/10.1109/IJCNN.2008.4633969, iSSN: 2161-4407
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (Dec 2015). https://doi.org/10.48550/arXiv.1512.03385, http://arxiv.org/abs/1512.03385, arXiv:1512.03385 [cs]
16. Hernandez, J., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: An Empirical Study of Oversampling and Undersampling for Instance Selection Methods on Imbalance Datasets. In: Ruiz-Shulcloper, J., Sanniti di Baja, G. (eds.) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. pp. 262–269. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (2013). https://doi.org/10.1007/978-3-642-41822-8_33
17. Houlsby, N., Huszár, F., Ghahramani, Z., Lengyel, M.: Bayesian Active Learning for Classification and Preference Learning (Dec 2011). https://doi.org/10.48550/arXiv.1112.5745, http://arxiv.org/abs/1112.5745, arXiv:1112.5745 [cs, stat]
18. Intelligence, I.: Drone technology uses and applications for commercial, industrial and military drones in 2021 and the future, https://www.businessinsider.com/drone-technology-uses-applications
19. Kaur, H., Pannu, H.S., Malhi, A.K.: A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions. ACM Computing Surveys **52**(4), 1–36 (Jul 2020). https://doi.org/10.1145/3343440, https://dl.acm.org/doi/10.1145/3343440
20. Kirsch, A., van Amersfoort, J., Gal, Y.: BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning (Oct 2019), http://arxiv.org/abs/1906.08158, arXiv:1906.08158 [cs, stat]
21. Krizhevsky, A.: Learning Multiple Layers of Features from Tiny Images p. 60
22. Kwolek, B., Koziarski, M., Bukala, A., Antosz, Z., Olborski, B., Wasowicz, P., Swadźba, J., Cyganek, B.: Breast Cancer Classification on Histopathological Images Affected by Data Imbalance Using Active Learning and Deep Convolutional Neural Network. In: Tetko, I.V., Kurkova, V., Karpov, P., Theis, F. (eds.) Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions, vol. 11731, pp. 299–312. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-30493-5_31, http://link.springer.com/10.1007/978-3-030-30493-5_31, series Title: Lecture Notes in Computer Science
23. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles (Nov 2017). https://doi.org/10.48550/arXiv.1612.01474, http://arxiv.org/abs/1612.01474, arXiv:1612.01474 [cs, stat]
24. Li, W., Li, J., Wang, Z., Polson, J., Sisk, A.E., Sajed, D.P., Speier, W., Arnold, C.W.: PathAL: An Active Learning Framework for Histopathology Image Analysis. IEEE Transactions on Medical Imaging **41**(5), 1176–1187 (May 2022). https://doi.org/10.1109/TMI.2021.3135002, conference Name: IEEE Transactions on Medical Imaging
25. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-Scale Long-Tailed Recognition in an Open World. pp. 2537–2546 (2019)
26. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (Jan 2019). https://doi.org/10.48550/arXiv.1711.05101, http://arxiv.org/abs/1711.05101, arXiv:1711.05101 [cs, math]

27. Mohammed, R., Rawashdeh, J., Abdullah, M.: Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. In: 2020 11th International Conference on Information and Communication Systems (ICICS). pp. 243–248 (Apr 2020). https://doi.org/10.1109/ICICS49469.2020.239556, iSSN: 2573-3346
28. Ren, R., Hung, T., Tan, K.C.: A Generic Deep-Learning-Based Approach for Automated Surface Inspection. IEEE Transactions on Cybernetics **48**(3), 929–940 (Mar 2018). https://doi.org/10.1109/TCYB.2017.2668395, conference Name: IEEE Transactions on Cybernetics
29. Richards, J.W., Starr, D.L., Brink, H., Miller, A.A., Bloom, J.S., Butler, N.R., James, J.B., Long, J.P., Rice, J.: ACTIVE LEARNING TO OVERCOME SAMPLE SELECTION BIAS: APPLICATION TO PHOTOMETRIC VARIABLE STAR CLASSIFICATION. The Astrophysical Journal **744**(2), 192 (Dec 2011). https://doi.org/10.1088/0004-637X/744/2/192, https://doi.org/10.1088/0004-637x/744/2/192, publisher: American Astronomical Society
30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. International Journal of Computer Vision **128**(2), 336–359 (Feb 2020). https://doi.org/10.1007/s11263-019-01228-7, http://arxiv.org/abs/1610.02391, arXiv:1610.02391 [cs]
31. Settles, B.: Active Learning Literature Survey. Technical Report, University of Wisconsin-Madison Department of Computer Sciences (2009), https://minds.wisconsin.edu/handle/1793/60660, accepted: 2012-03-15T17:23:56Z
32. Shannon, C.E.: A Mathematical Theory of Communication p. 55
33. Shi, X., Dou, Q., Xue, C., Qin, J., Chen, H., Heng, P.A.: An Active Learning Approach for Reducing Annotation Cost in Skin Lesion Analysis. In: Suk, H.I., Liu, M., Yan, P., Lian, C. (eds.) Machine Learning in Medical Imaging. pp. 628–636. Lecture Notes in Computer Science, Springer International Publishing, Cham (2019)
34. Wang, X., Liu, B., Cao, S., Jing, L., Yu, J.: Important sampling based active learning for imbalance classification. Science China Information Sciences **63**(8), 182104 (Aug 2020). https://doi.org/10.1007/s11432-019-2771-0, http://link.springer.com/10.1007/s11432-019-2771-0
35. Wightman, R., Soare, A., Arora, A., Ha, C., Reich, C., Raw, N., Kaczmarzyk, J., MrT23, Mike, SeeFun, Contrastive, Rizin, M., Kim, H., Kertész, C., Mehta, D., Cucurull, G., Singh, K., Han, Tatsunami, Y., Lavin, A., Zhuang, J., Hollemans, M., Sameni, S., Shults, V., Wang, X., Kwon, Y., Uchida, Y., Zhong, Z., Comar: rwightman/pytorch-image-models: v0.6.5 Release (Jul 2022). https://doi.org/10.5281/ZENODO.4414861, https://zenodo.org/record/4414861
36. Yin, X., Chen, Y., Bouferguene, A., Zaman, H., Al-Hussein, M., Kurach, L.: A deep learning-based framework for an automated defect detection system for sewer pipes. Automation in Construction **109**, 102967 (Jan 2020). https://doi.org/10.1016/j.autcon.2019.102967, https://www.sciencedirect.com/science/article/pii/S0926580519307411
37. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features (May 2019), https://arxiv.org/abs/1905.04899v2
38. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond Empirical Risk Minimization. arXiv:1710.09412 [cs, stat] (Apr 2018), http://arxiv.org/abs/1710.09412, arXiv: 1710.09412

39. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep Long-Tailed Learning: A Survey (Oct 2021), http://arxiv.org/abs/2110.04596, arXiv:2110.04596 [cs]