ce/papers Proceedings in civil engineering

Ernst & Sohn A Wiley Brand

ORIGINAL ARTICLE

EUROSTRUCT

metal-dacl: Image-Based Automated Damage Recognition for Steel Bridge Inspections

Johannes Flotzinger[1] | Diego Mediel-Cuadra[1] | Jonas Zausinger[2] | Fabian Deuser[1] | Lukas Rauch[1] | Thomas Braml[1]

Johannes Flotzinger
University of the Bundeswehr Munich
Werner-Heisenberg-Weg 39
85579 Neubiberg
Email: johannes.flotzinger@unibw

[1] UniBw, Neubiberg, Germany
[2] TUM, Munich, Germany

Abstract

As infrastructure ages and the number of structures requiring inspection increases, effective monitoring of damage in built structures has become more crucial than ever. Staff shortages and budget constraints can make it difficult for authorities to conduct the necessary frequent inspections. To address these challenges, companies and research institutions are increasingly exploring digital approaches to building inspection. Digitalized inspection processes involve creating a digital shadow of the structure, which combines a BIM model with a record of classified, measured, localized and assessed defects. Key to this approach is the deployment of transformer-based architectures for image-based automated damage recognition. Accurate defect recognition is essential for evaluating the condition of specific areas and assessing the overall integrity of a structure. This paper presents a dataset extension for the dacl10k dataset tailored to steel defect recognition on bridges. We manually assigned polygonal annotations to 3,737 images of dacl10k that showed steel bridges or building parts. Despite the challenging nature of this segmentation dataset extension, our baseline model achieves a mean Intersection-over-Union of 17.37%. This result provides a valuable reference point for future models and highlights the complexity inherent in detecting fine-grained steel defects. We conduct a detailed analysis to uncover the factors limiting model performance and suggest pathways for improvement.

Keywords

Bridge Inspection, Steel Defect Recognition, Deep Learning, Computer Vision

1 Introduction

The structural integrity of transportation infrastructure, particularly bridges, is a critical concern for public safety and economic stability. According to recent governmental reports, nearly one in three bridges in the United States requires repair or replacement [1]. As illustrated in Figure 1, the number of bridges classified as needing replacement or rehabilitation is constantly high, reaching approximately 222,000 by 2024. This alarming statistic underscores the urgent need for more efficient and scalable inspection methodologies that can address both the growing volume of infrastructure and the limitations of manual inspection practices.

Traditional bridge inspections are labour-intensive, time-consuming, and susceptible to human error. Furthermore, a shortage of skilled inspectors and constrained public maintenance budgets compound the difficulty of maintaining structural safety. In response, both industry and academia have intensified efforts to develop automated, image-based inspection technologies powered by computer vision and deep learning. These technologies offer the potential to enhance inspection speed, consistency, and objectivity, making them particularly valuable for large-scale infrastructure monitoring.

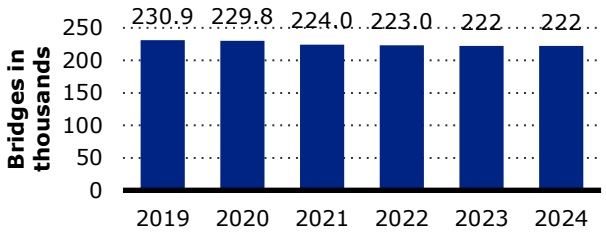


Figure 1 Number of bridges in need of replacement or rehabilitation in the United States from 2019 to 2024 (in 1,000s) [1].

One significant contribution to the advancement of automated inspection is the dacl10k dataset, which currently

represents the most diverse and extensive dataset for automated concrete bridge inspections [2]. dacl10k includes 9,920 annotated images and supports pixel-level classification of 13 distinct defect classes based on German inspection standards and six bridge components. Despite its benchmark character, the dataset primarily focuses on concrete defects, leaving a gap in resources for steel defect recognition — a critical aspect in the assessment of steel and steel-composite bridges.

To address this shortfall, we propose metal-dacl, an extension of the dacl10k dataset specifically targeting steel defect segmentation. This work expands the dataset with 3,737 manually annotated steel defect images with pixel-level segmentation masks across 10 steel defect classes, following established German inspection standards[3, 4]. To evaluate the feasibility of automated defect detection, we introduce a baseline model - a vision transformer (MaxViT) [5] with a Feature Pyramid Network (FPN) decoder [6] - and report per-class IoU across the spectrum of defects.

This research aims to advance the development of tools for automatic, automated, and autonomous bridge inspections. The models trained on our newly introduced dataset extension are designed for deployment in mobile and drone-based inspection systems, significantly improving the accuracy, efficiency, and consistency of damage detection, documentation, and assessment. Our contributions are:

- We release metal-dacl, the first dataset extension with polygonal annotations for defect segmentation on steel bridges.
- We provide a vision transformer baseline and comprehensive performance analysis across defect types.
- We perform an in-depth analysis of data-specific challenges to surface defect ambiguity, class imbalance, and inconsistent annotation semantics.

2 Dataset

The dacl10k dataset comprises 9,920 images collected during bridge inspections [2]. Each image is annotated at pixel level using polygonal segmentation, enabling semantic differentiation between various defect types and structural components. The dataset includes 19 distinct classes, of which 13 are damage classes such as cracks, spalling, rust, and efflorescence, while the remaining 6 correspond to building components including bearings and drainage systems. The classification and annotation process adheres to the guidelines of the German bridge inspection standard DIN 1076 [3] and the RI-EBW-PRÜF [4] directive. All annotations are manually generated to ensure alignment with inspection protocols. The dataset is designed to support research in image-based damage detection and classification, particularly in the context of concrete bridge structures. Its scope addresses the requirement for annotated visual data in the development and validation of deep learning models applied to infrastructure inspection.

Several datasets have been developed to support research on image-based detection of surface anomalies in steel structures. The NEU Surface Defect Dataset is one of the earliest and most widely used benchmarks [7]. It contains 1,800 grayscale images of six typical surface defect types found on hot-rolled steel strips. These include rolled-in scale, patches, crazing, pitted surface, inclusions, and scratches. Each image is labeled by defect class, and the dataset is frequently used to evaluate classification and detection models.

The Severstal Steel Defect Detection dataset, released as part of a Kaggle competition, includes approximately 12,568 high-resolution images of steel sheets. Defects are annotated at the pixel level and categorized into four types. The dataset includes both defective and defect-free samples, providing a basis for evaluating segmentation performance in production-like settings [8].

The GC10-DET dataset extends the scope of defect types by including ten categories such as punching, inclusion, oil spots, and welding faults. It consists of 3,570 grayscale images with pixel-level annotations and has been used to evaluate more complex models in multi-class scenarios [9]. In addition, the FSC-20 dataset aggregates 20 steel surface defect types from multiple sources, supporting research in few-shot learning under limited data conditions [10].

More recently, the VISION benchmark, presented at the CVPR 2023 Vision-Based Industrial Inspection workshop, consolidated 14 inspection datasets across 44 defect types and 18,000 annotated images [11]. While not limited to steel, this benchmark includes steel surface defects and standardizes annotation formats across datasets. It reflects a broader effort to establish unified benchmarks in industrial inspection tasks.

These datasets provide valuable resources for developing and evaluating defect detection models. However, their focus is primarily on flat steel sheets, and they offer limited representation of bridge-related steel defects. To address this gap, we introduce the metal-dacl extension as an extension of dacl10k, tailored to structural steel components in bridge environments.

Through metal-dacl we extend the dacl10k dataset with polygonal annotations for enabling image-based automated damage recognition on steel structures. In total 3,737 images were labeled with at least one steel defect. Most of these images show steel bridges, building parts of concrete bridges made of steel, or traffic-sign bridges. Table 1 gives an overview of these classes. There are two classes that represent components: *screw* and *weld,* as well as eight steel defects*.*

3 Experiments

The objective of this study is to train and analyze a segmentation baseline for automated steel defect recognition using the metal-dacl extension. The model architecture integrates two components: MaxViT-Base as image encoder and Feature Pyramid Network (FPN) which represents the segmentation architecture. MaxViT-Base is a vision transformer that combines local and global attention mechanisms through a multi-axis attention design. It processes image features across multiple scales using a hybrid of

convolution and Transformer operations, which allows to model spatial dependencies with controlled computational cost [5]. The Feature Pyramid Network (FPN) operates as a feature aggregation module. It constructs a pyramid by combining low-level and high-level features through a combination of bottom-up processing and top-down fusion with lateral connections. This configuration enables the model to retain detailed spatial information while capturing broader semantic context across image scales [6].

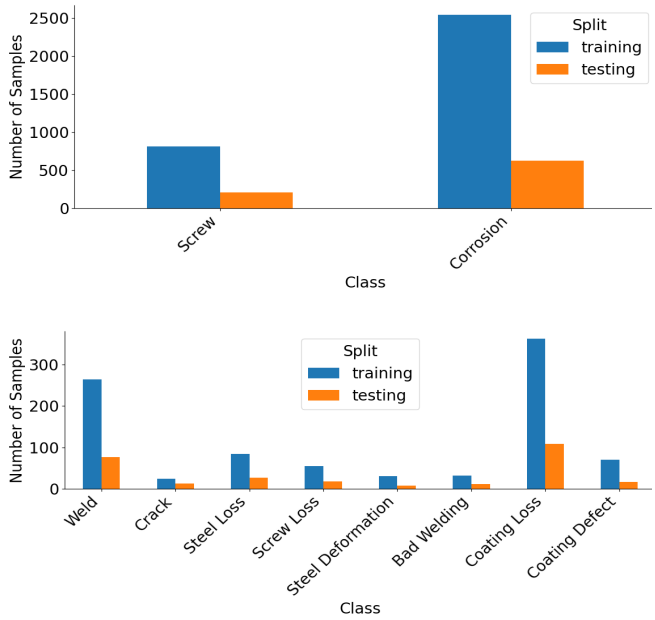


Figure 2 Sample-level class distribution in the metal-dacl extension. The histograms show the number of annotated instances per defect class, separated by training and testing splits. The upper plot displays the two most frequent classes, while the lower plot presents the remaining eight defect classes.

All input images are resized to 512×512 pixels. Data augmentation is applied during training, including random rotation and flipping, to increase variability in the training data. The training procedure consists of 50 epochs using a supervised learning framework.

3.1 Metrics

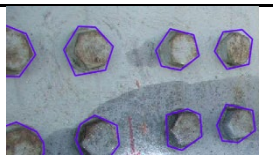
The evaluation of segmentation performance in this study is based on the Intersection-over-Union (IoU) metric. IoU is a standard metric for assessing the accuracy of predicted segmentation masks relative to ground truth annotations in pixel-level classification tasks.

Mathematically, the IoU for a single class is defined as:

$$IoU = \frac{|A \cap B|}{|A \cup B|}$$

Where A is the set of predicted pixels, B is the set of ground truth pixels, $|A \cap B|$ denotes the set of pixels correctly predicted as class members (true positives) and $|A \cup B|$ the union of predicted and ground truth pixels, which includes true positives, false positives, and false negatives. Thus, the IoU measures the proportion of overlap between the predicted and annotated regions for a given class.

Table 1 Class guideline of steel defects added to dacl10k

Steel Damage Class	Description	Sample
Screw	-Screws/bolts and rivets	
Weld	-All weld types	
Corrosion	-Can appear on Steel and all Components made of steel	
Crack	-Crack in the steel structure, welds, srews/bolts or rivets	
Steel Loss	- Parts of the affected steel component are completely missing (holes), most likely due to heavy corrosion	
Screw Loss	- Screws/bolts or rivets are missing	
Coating Loss	- Results from loss of adhesion of the CS/paint film - Subjacent steel surface is visible	
Steel Deformation	- Steel/building part seems deformated - Subtypes of deformation: Buckling, deflection, distortion, and torsion	
Bad Welding	- Welding spatters - Undercut - Concavities - Porous	
Coating Defect	- Blistering - Flaking	

This metric yields a value between 0 and 1, where 1 indicates perfect alignment between the predicted and ground

truth masks, and 0 indicates no overlap. IoU is computed individually for each defect class, and the overall performance is expressed as the mean IoU across all classes.

In the context of this study, IoU is used to quantify the extent to which the segmentation model correctly identifies and localizes defect regions on steel bridge components based on the pixel-wise annotations in the metal-dacl extension. To enhance readability, IoU scores in Table 2 are presented as percentages.

3.2 Results

The evaluation of our semantic segmentation baseline on the extended steel-defect dataset (Table 2) reveals substantial variability across defect classes, pointing to both the strengths and limitations of the underlying initial version of metal-dacl.

The baseline achieves a mean IoU of 17.37%, establishing a solid foundation for continued research in semantic segmentation of steel defects. While further refinements are necessary to reach deployment-ready performance, this result affirms the overall feasibility of our annotation scheme and model setup in a complex, real-world scenario. Notably, key defect classes yield comparably strong results to those reported on the dacl10k dataset[3, 4]. In that work, the top-performing model reached 45% IoU on the "Ruststain" (Rust) class, with the Engineer vs. Machine study demonstrating that this level of accuracy already surpasses manual expert annotations.

Among the individual classes, Corrosion stands out with the highest IoU of 50%. This superior performance is likely due to its strong presence and consistency in appearance across the dataset — indeed, *Corrosion* patches are the most representative and frequently occurring defects in our annotated images. Similarly, *Screw* shows reliable segmentation performance (IoU = 38%), which can be attributed to their distinctive, well-defined edges and the fact that the *Screw* class constitute the second largest class in terms of sample count. The clear, high-contrast visual cues associated with screws allow the network to learn robust feature representations despite the overall dataset's complexity.

The network also performs moderately well on *Weld* (IoU = 26%), indicating that weld seams — though sometimes visually similar to surrounding steel — nonetheless possess enough structural consistency for the model to distinguish them with reasonable accuracy. *Coating Loss* defects (IoU = 21%) fall into the upper mid-range of performance. However, this result is tempered by the fact that it encompasses a broad spectrum of visual patterns; further discussion of this issue is provided in the Limitations chapter.

By contrast, several defect classes exhibit near-zero or very low IoU scores, highlighting fundamental challenges in capturing their subtle characteristics. *Crack* and *Steel Deformation* defects both register an IoU of 0%, implying that the model essentially fails to detect any true positive pixels for these categories. Similarly, *Coating Defect* is nearly ignored (IoU = 1.5%), suggesting that the intricate, fine-grained visual signatures of small surface blemishes elude the current architecture. The reason for these low IoU values is that the dataset either does not contain enough examples of these defects or that the visual difference between defect and non-defect is too subtle for the base model to capture without special training strategies or higher resolution images. In the next section we will analyse this in more detail and outline strategies for dealing with these problems.

Defect classes with intermediate-low performance—namely *Steel Loss* (IoU = 13%) and *Screw Loss* (IoU = 11%)—fall into what can be described as the lower mid-field. Both steel-loss categories involve irregular, often irregularly shaped voids or missing material that can blend into the surrounding structures, thus posing difficulties for the segmentation network.

Taken together, these results underscore that *Corrosion*, *Screw*, and *Weld* are currently learnable with a degree of accuracy. The remaining classes — particularly *Crack*, *Steel Deformation*, and *Coating Defect* — prove almost intractable under our baseline setup. In practical terms, a mean IoU of 17% is insufficient for real-world bridge-inspection scenarios, where false negatives or misclassifications of subtle defects could have serious safety implications. The subsequent analysis section delves into the root causes of these performance gaps and outlines potential avenues for improving defect detectability, such as augmenting dataset size, enhancing annotation consistency, and exploring more advanced network architectures.

Table 2 Intersection over Union scores per class

Class Label	IoU (%)
Screw	37.72
Weld	26.17
Corrosion	49.95
Crack	0.00
Steel Loss	13.40
Screw Loss	10.82
Steel Deformation	0.00
Bad Welding	13.42
Coating Loss	20.80
Coating Defect	1.46
Mean	17.37

4 Limitations

Table 3 presents three representative samples from the test set alongside our model's corresponding predictions. These qualitative examples, taken together with the quantitative results in Table 2, highlight several key limitations of our current approach:

First, *Screw*, *Weld*, and *Corrosion* defects are segmented reasonably well (see Sample #1 to #3 in Table 3). The model consistently identifies and outlines *Screws* and *Weld* seams with relative accuracy, and *Corrosion* patches are detected robustly. However, not all classes show such performance.

The *Crack* class suffers from severe underrepresentation: only 25 cracked-weld examples appear in the training split and a mere 13 in the test split (see Figure 2). Consequently, the model has almost no opportunity to learn crack-specific features, resulting in zero IoU. As shown in Table 1, the class is finely grained and contains a small number of pixels. This makes it challenging for the network to learn meaningful representations and achieve accurate segmentation. More annotated *Crack* samples are needed before the network can reliably distinguish these narrow fissures from their surroundings.

Similarly, *Steel Loss* proves to be a particularly complex category. In our annotations, *Steel Loss* includes both holes that have corroded entirely through a steel member and areas where surface steel has begun flaking away. In the case of holes, the background visible through the opening can be anything — from concrete to the sky — forcing the model to rely heavily on contextual cues rather than consistent visual texture. This variation in context makes it difficult for the network to learn a unified concept of *Steel Loss*.

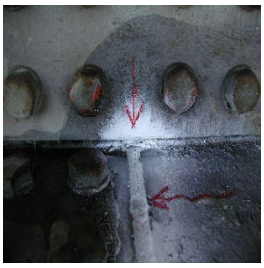
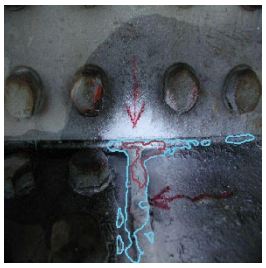
The *Screw Loss* category is also data-scarce, with just 73 examples across the train and test splits, and it is inherently linked to *Steel Loss* when *Screws* have corroded away. Because *Screw Loss* annotations often overlap visually with the broader *Steel Loss* patterns, the model struggles to disentangle one from the other, leading to low segmentation accuracy.

For *Steel Deformation*, most samples in our dataset occur on secondary elements such as drainage pipes or signage. There are very few instances of deformation on primary structural components like girders. As a result, the network has not been exposed to — and therefore cannot generalize to — cases where load-induced bends or buckles appear on major load-bearing members. A stronger dataset of deformed girders and beams is needed.

The *Bad Welding* class (sample #2 in Table 3) further illustrates annotation ambiguity. Although most visibly cracked *Weld* seams are annotated as *Bad Welding*, the underlying cause of the defect (e.g., improper welding technique versus subsequent cracking) is seldom evident in the image. In practice, many *Bad Welding* annotations simply mark the *Crack* itself, effectively overlapping with the *Crack* class. Because our model sees predominantly white areas (i.e., crack-detection spray) for bad welds, it confuses these regions with *Crack* annotations, thereby hampering its ability to learn a distinctive *Bad Welding* concept. Moreover, the limited number of bad-weld examples means the network rarely encounters enough variety to form a robust representation.

Coating Loss appears to be a mid-range performer by metric alone (see Table 2, IoU = 21%), but a close qualitative examination (sample #3 in Table 3) reveals significant semantic overlap with *Corrosion*. Many areas labeled as *Coating Loss* also exhibit early stages of *Corrosion*, and our annotation guidelines permitted dual labeling (*Coating Loss* + *Corrosion*). In practice, however, if *Corrosion* is present, it should not be overlaid with an additional *Coating Loss* label — doing so confuses the model into believing that any corroded pixel must also be a site of *Coating Loss*. Ensuring mutual exclusivity between these two classes within the annotation protocol is essential for optimizing model performance.

Table 3 Three samples drawn from test set and prediction of our model.

#	Raw	Prediction
1		
2		
3		

Finally, *Coating Defect* — which encompasses phenomena such as paint flaking or blistering — requires more samples and possibly a finer-grained breakdown. Flaking and blistering present visually distinct patterns (sharp, irregular edges for flakes versus smoother, localized raised areas for blisters), yet we currently group them together. This heterogeneity likely contributes to the model's very low IoU for *Coating Defects* (1.5%). Separating these subtypes into flaking and blistering classes may help the network learn each defect's unique visual signature.

In summary, while some defect classes (*Screw*, *Weld*, and *Corrosion*) are relatively well learned, many other categories suffer from extreme class imbalance, inconsistent annotation guidelines, and overlapping semantics. Addressing these limitations — by expanding the dataset with underrepresented categories, refining labeling rules (especially for *Coating Loss* versus corrosion and *Bad Welding* versus *Crack*), and possibly subdividing heterogeneous classes — will be essential for improving the overall segmentation performance.

5 Conclusion

Our analysis indicates that, while the segmentation baseline demonstrates promising results on a handful of defect classes, significant improvements are required before this approach can be deployed in practice. In particular, defects such as *Corrosion*, *Screw*, and *Weld* achieve the strongest performance — reflecting their relatively consistent visual appearance and adequate representation in the dataset. In contrast, some classes present challenges due to either ambiguous annotation guidelines or limited representation in the dataset. For instance, *Coating Loss* labels often overlap with corrosion regions, suggesting a need for more consistent annotation practices to better distinguish between these defects. Additionally, categories such as *Bad Welding*, *Steel Deformation*, *Crack*, and *Screw Loss* currently have relatively few annotated examples, making it difficult to train models with high reliability for these cases.

In summary, our work highlights two key findings for future research. First, we must substantially expand the dataset — with particular emphasis on rare or complex defect types — and rework the annotation guidelines to eliminate semantic overlap. Second, we should explore more efficient labeling paradigms (e.g., "weakly supervised" [12] or "lazy student" approaches [13]) to accelerate the creation of high-quality ground truth masks without incurring prohibitive manual-annotation costs. Only by combining richer and more consistent annotations with scalable labeling methodologies can we hope to push mean IoU values beyond the current 17% ceiling and move toward reliable, automated steel-defect segmentation for bridge inspection.

References

[1] American Road & Transportation Builders Association (ARTBA). (2024). *ARTBA-Bridge-Report*

[2] Flotzinger, J.; Rösch, P. J.; Braml, T. (2024). dacl10k: Benchmark for Semantic Bridge Damage Segmentation, *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*Presented at the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), , IEEE, Waikoloa, HI, USA, 8611–8620. doi:10.1109/WACV57701.2024.00843

[3] DIN 1076:2024-02, Ingenieurbauwerke im Zuge von Straßen und Wegen_- Überwachung und Prüfung. (n.d.), DIN Media GmbH. doi:10.31030/3510975

[4] Guideline for the uniform acquisition, assessment, recording and evaluation of results of structural inspections (RI-EBW-PRÜF). (2017, February)

[5] Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; Li, Y. (2022, September 9). MaxViT: Multi-Axis Vision Transformer, arXiv. doi:10.48550/arXiv.2204.01697

[6] Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. (2017, April 19). Feature Pyramid Networks for Object Detection, arXiv. doi:10.48550/arXiv.1612.03144

[7] Song, K.; Yan, Y. (2013). A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects, *Applied Surface Science*, Vol. 285, 858–864. doi:10.1016/j.apsusc.2013.09.002

[8] Alexey Grishin and BorisV and iBardintsev and inversion and Oleg. (2019). Severstal: Steel Defect Detection, from https://kaggle.com/severstal-steel-defect-detection, accessed 31-5-2025

[9] Lv, X.; Duan, F.; Jiang, J.; Fu, X.; Gan, L. (2020). Deep Metallic Surface Defect Detection: The New Benchmark and Detection Network, *Sensors*, Vol. 20, No. 6, 1562. doi:10.3390/s20061562

[10] Zhao, W.; Song, K.; Wang, Y.; Liang, S.; Yan, Y. (2023). FaNet: Feature-aware network for few shot classification of strip steel surface defects, *Measurement*, Vol. 208, 112446. doi:10.1016/j.measurement.2023.112446

[11] Bai, H.; Mou, S.; Likhomanenko, T.; Cinbis, R. G.; Tuzel, O.; Huang, P.; Shan, J.; Shi, J.; Cao, M. (2023, June 18). VISION Datasets: A Benchmark for Vision-based InduStrial InspectiON, arXiv. doi:10.48550/arXiv.2306.07890

[12] Yang, X.; Gong, X. (2023). Foundation Model Assisted Weakly Supervised Semantic Segmentation, arXiv. doi:10.48550/ARXIV.2312.03585

[13] Ke, R.; Bugeau, A.; Papadakis, N.; Schuetz, P.; Schönlieb, C.-B. (2019). Learning to segment microscopy images with lazy labels, arXiv. doi:10.48550/ARXIV.1906.12177