

Image-based Detection of Structural Defects using Hierarchical Multi-Scale Attention*

Christian Benz^[0000-0001-9915-0057] and Volker Rodehorst^[0000-0002-4815-0118]

Bauhaus-Universität, Weimar, Germany
{christian.benz, volker.rodehorst}@uni-weimar.de

Abstract. With improving acquisition technologies, the inspection and monitoring of structures has become a field of application for deep learning. While other research focuses on the design of neural network architectures, this work points out the applicability of transfer learning for detecting cracks and other structural defects. Being a high-performer on the Cityscapes benchmark, *hierarchical multi-scale attention* [43] also renders suitable for transfer learning in the domain of structural defects. Using the joint scales of 0.25, 0.5, and 1.0, the approach achieves 92% mean intersection-over-union on the test set. The effectiveness of multi-scale attention is demonstrated for class demarcation on large scales and class determination on lower scales. Furthermore, a *line-based tolerant intersection-over-union* metric is introduced for more robust benchmarking in the field of crack detection. The dataset of 743 images covering *crack*, *spalling*, *corrosion*, *efflorescence*, *vegetation*, and *control point* is unprecedented in terms of quantity and realism.

Keywords: Deep learning · Structural defects · Crack detection · Hierarchical multi-scale attention.

1 Introduction

The field of structural health monitoring (SHM) deals with the regular inspection and assessment of engineering structures, such as bridges, to ensure their safe use. With the ongoing digitalization in SHM, the amount and quality of imagery of critical infrastructure is successively growing. Automated image-based detection of structural defects can substantially support the human decision makers in assessing the operability of a structure. An appropriate set of images can serve several purposes, including the 3D reconstruction and maintenance of a digital twin of the structure. Furthermore, high-quality imagery can effectively

* This version of the contribution has been accepted for publication, after peer review (when applicable) but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/978-3-031-16788-1_21. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use <https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms>.

be used for non-destructive SHM. One informative surface property for condition assessment is the formation of cracks. Other defects, however, also provide substantial insights, defects such as spalling, corrosion, efflorescence, and vegetation. Furthermore, for purposes of georeferencing, control points constitute another worthwhile class for automated detection.

Unlike other work in the domain, that dedicates to the design of suitable artificial neural network architectures, the here presented work applies transfer learning. For that purpose, a high-ranking approach in the Cityscapes benchmark [10] with accessible code is used, the hierarchical multi-scale attention (HMA) approach by [43]. The effectiveness of HMA for semantic segmentation of structural defects is demonstrated and the role of multi-scale attention further investigated. For all classes except spalling, larger scales seem to draw more attention towards class demarcation, while smaller scales rather engage in content determination.

The plausible evaluation of detection performance is key to advancements in the field. The appropriateness of area-based metrics, such as F_1 score and intersection-over-union (IoU), for crack detection are debatable. Cracks can conceptually rather be considered as lines, thus, the here introduced metric of line-based tolerant IoU can support more robust benchmarking.

The contributions of this work are fourfold: (1) The creation of a dataset¹ that is unprecedented with respect to realism and classes. (2) A demonstration of the applicability of transfer learning to a state-of-the-art semantic segmentation approach for the domain of structural defects². (3) An analysis of the role of attention in multi-scale fusion of different classes. And (4) the presentation of the line-based tolerant IoU evaluation metric that is more appropriate for crack segmentation.

2 Related work

The related work covers the detection of structural defects resp. anomalies, which include cracks, and the field of transfer learning.

2.1 Anomaly Detection

In recent years, the emergence and utilization of data-driven methods in image processing as well as the potential impact in SHM has fueled the research activity in image-based anomaly detection. In the context of structural health monitoring, the term *anomaly* refers to irregularities in the structure that potentially impede its functionality. Here it is interchangeably used with (*structural defects*).

Cracks form one prominent class of defects. Due to the high relevance for SHM, the attention on visual crack detection has steadily been growing [32].

¹ The dataset is available at <https://github.com/ben-z-original/s2ds>.

² Code is available at <https://github.com/ben-z-original/detectionhma>

Especially approaches based on convolutional neural networks (CNN) have had a large impact on the field of image-based object detection [11,26,41,42,22] and are increasingly used for crack detection. Image classification is one conceptual approach to crack detection, where each patch of the image receives a positive or negative response. [13] provides a comparatively large dataset, SDNET, and e.g. [12,48,37] propose image classification-based CNNs for crack detection. With the introduction of fully-convolutional neural networks (FCN) [31], semantic segmentation – producing dense, pixelwise predictions – have become the natural approach for many applications, including crack detection: The more precise localization facilitates further geometric analysis such as crack width estimation, cf. [3].

Among the fully-convolutional approaches to crack detection are [46,45,30,54] [29,15]. [30] propose a network topology based on U-NET [40] and demonstrate its superiority over a simpler FCN design. U-NET makes use of skip connections between encoder and decoder to convey information from earlier stages to later stages of the model. Conceptually similar but more elaborated topologies are used by [54,29], who independently introduce a separate fusion stage: Skip connections do not map from encoder to decoder, but to a separate fusion module. This module, a cascade of convolutional layers, is fed from different scales of encoder and decoder and, through upsampling and fusion, derives a prediction.

With transformer approaches gaining more relevance in visual applications [14,6,51], they have also been explored for crack detection. [28] extend the Seg-Net approach [2] by self-attention modules in order to exploit long-range dependencies of cracks. While the convolutional mechanism is meant to capture fine details of thin cracks, the attention module is designed to form a continuous crack representation.

Providing substantial insights into the health of a structure, the image-based detection of structural defects, other than cracks, has also gained attention. Classification-based detectors for corrosion and spalling are proposed, by e.g. [1,34,39,17,18], while others [33,25,16,23,24] choose segmentation approaches at pixel-level. [25,16] for instance, use, investigate, and show the applicability of the U-NET for corrosion detection, even though inconsistently outperformed by the FCN. [33] create a synthetic dataset and achieve IoUs of roughly 40% for exposed reinforcement bar (similar to corrosion) and concrete damage on a very small real-world test set. Concrete damage and its severity are also targeted by [36], who use a bounding box-based detection approach to localize defects. Being also applied by [7], the bounding box approach, however, appears inappropriate for less compact defects such as cracks. [38] use mold, stain, and deterioration as classes and retrain VGG [41] for classification.

2.2 Transfer Learning

While most of the related work engages in designing a network architecture, customized for anomaly detection, transfer learning is rather exceptionally applied. Transfer learning refers to a learning process, where the learner must perform

multiple different, but related tasks [20]. In the context of artificial neural networks, it typically refers to a change in targets, such as the visual categories to be classified. The underlying assumption is that different targets share low-level features and, thus, can mutually benefit. The effectiveness of transfer learning has been demonstrated for various tasks, including domain adaptation and learning from little data.

The success of transfer learning being one reason for most of the deep learning libraries to provide a model zoo, a collection of established deep learning models, such as [22,8,9]. The DeepLab approach [8,9] uses an encoder-decoder design and atrous spatial pyramid pooling (ASPP), to efficiently extend the receptive field. An effective source for performance comparison is provided through benchmarking challenges such as Cityscapes [10], KITTI [19], COCO [27], or ADE20K [52]. At the time of experimentation, hierarchical multi-scale attention (HMA) [43] was the best performing model in the Cityscapes semantic segmentation challenge with accessible code. It was, thus, considered to be a powerful approach for a domain adaptation to structural defects through transfer learning.

3 Data

Even though growing, the number of publicly available datasets for structural defects is rather limited. For cracks a number of datasets are available, such as [53,54,29,13]. They differ in annotation style (image-, line-, segmentwise), represented surfaces (asphalt, concrete, stone, etc.), and level of difficulty (presence of crack-resembling artifacts). Datasets for structural defects other than cracks are less common or incomplete, e.g. [35,4]. Potential reasons are low accessibility of defects on structures, vagueness of defect boundaries, high variance of surfaces and structures, high annotation effort (involving experts), and commercially induced reluctance to data publication. Due to lack of data, [33] created a synthetic dataset.

Table 1: Overview of the structural defects dataset (S2DS).

Class	Training			Validation			Test		
	Images	Pixels	Area	Images	Pixels	Area	Images	Pixels	Area
Background	556	519.9 M	88.1 %	87	83.8 M	91.9 %	93	87.7 M	90.0 %
Crack	180	0.6 M	0.1 %	25	0.1 M	0.1 %	27	0.1 M	0.1 %
Spalling	151	39.2 M	6.6 %	23	3.3 M	3.6 %	20	4.2 M	4.3 %
Corrosion	209	8.8 M	1.5 %	36	0.5 M	0.6 %	38	0.9 M	0.9 %
Efflorescence	96	4.6 M	0.8 %	13	0.6 M	0.7 %	17	1.5 M	1.5 %
Vegetation	97	15.7 M	2.7 %	16	2.7 M	2.9 %	18	2.9 M	2.9 %
Control Point	70	1.5 M	0.3 %	9	0.2 M	0.2 %	10	0.2 M	0.3 %
Total	563	590.35 M	100 %	87	91.23 M	100 %	93	97.52 M	100 %

Scarcity and inappropriateness of the available datasets rendered necessary the creation of a suitable dataset, the *structural defect dataset*, S2DS. For that purpose, 743 patches of size 1024×1024 px were extracted from 8,435 images taken by structural inspectors at real inspection sites. The images were acquired with various different camera platforms, such as DSLR cameras, mobile phones, or drones (UAS). The quality and resolution of many of the images were insufficient for usage due to invisibility of defects or severe blur. A considerable number of selected patches, however, still vary in quality, i.e. in sharpness, lighting conditions, and color constancy. The images were selected and labeled by one trained computer scientist. For highest diligence and accuracy in labeling finest cracks, the scaling and blending options in the available annotation tools were too limited. These limitations and the comparatively low number of images rendered GIMP a suitable tool for annotation. Table 1 provides an overview of the dataset. Figure 3 and 4 in the results section convey a visual impression of the dataset. The dataset was manually split into subsets of 75% for training, 12% for validation, and 13% for testing. In order to get a realistic assessment given proper image material, only images with a fair chance of recognition made their way into the test set. For the other subsets, however, images with blurry, poorly resolved, or hardly visible defects were considered to enrich the yet relatively small training set. Due to the nature of the classes, the dataset is highly unbalanced with respect to the number of pixels and area per class. The imbalance in the number images is due to the imagery provided by structural inspectors: the prevalence of cracks, spalling, and corrosion as well as their major relevance for structure inspection, lead to higher amounts of image material of these classes. The imbalance in the number of pixels indicates the global underrepresentation of the crack class, which is, furthermore, confirmed by the relative occupation in terms of area.

The selected portfolio of classes contains crack, spalling, corrosion, efflorescence, vegetation, and control point: *cracks* represent linear fractures in the material, *spalling* refers to a material detachment from the surface, *corrosion* denotes the rust formation by oxidizing metal parts, *efflorescence* are depositions of dissolved chemicals on the structure’s surface, *vegetation* refers to surficial plant growth, and *control points* are geodetic fiducial markers for georeferencing. Control points do not form a class of structural defects. They are, however, substantial for georeferencing and SHM and, thus, are included and, for simplicity, referred to as structural defects in this work.

4 Hierarchical Multi-Scale Attention

At the time of experimentation HMA [43], was the highest ranked approach with publicly accessible code in the pixel-level semantic segmentation benchmark of Cityscapes [10]. It has recently been surpassed by [5], who introduce a structured boundary-aware loss. Applying this loss to HMA, an improvement of 0.5% points was achieved on the benchmark. As of May 2022, HMA occupies the second place of approaches with published code and the tenth place in the

overall competition. It, thus, can still be considered a state-of-the-art approach to semantic segmentation.

CNNs often struggle with the detection of objects that occur in various sizes [44]. To incorporate multiple scales, HMA proposes a dynamic combination of results from different scales based on simultaneously generated attention maps. The attention maps are contrastively learned based on two scales only. For inference, however, the number of scales can be arbitrarily chosen.

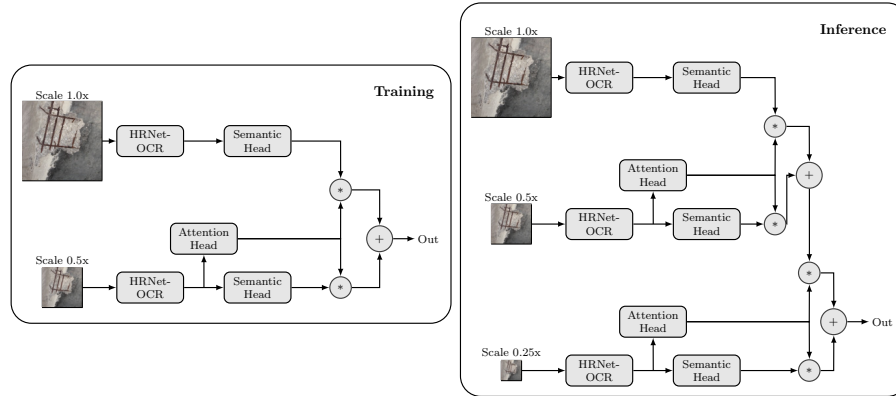


Fig. 1: Hierarchical multi-scale attention (HMA). Based on [43].

Figure 1 provides a high-level overview on the training and inference procedure. During training, the input image passes through the backbone (HRNet-OCR) [47] and the head for semantic segmentation in two different scales. On the smaller scale, the image additionally runs through the attention head. The attention head returns high values for image regions that obtain high attention for the given scale. For low-attention regions the results from the other scale gain higher relevance.

A recursive application of contrastive attention allows for arbitrarily many scales in inference, cf. Figure 1 (bottom). The results from the two larger scales are again merged, weighted by the pixel-wise attention. These weighted results are again weighted and fused based on the attention maps from the next scale. The scales shown in Figure 1 represent the scales actually used for training and inference for structural defects.

HMA uses the HRNet-OCR [47] as backbone, where OCR refers to object-contextual representations. [47] splits the image into regions, for which a region representation is computed by aggregation of pixel representations. Based on the relation of pixels and regions, an object-contextual representation is derived, that augments the pixel’s representation. Each pixel, thereby, obtains more information about its context.

The fully-convolutional head used for semantic segmentation consists of $(3 \times 3 \text{ conv}) \rightarrow (\text{BN}) \rightarrow (\text{ReLU}) \rightarrow (3 \times 3 \text{ conv}) \rightarrow (\text{BN}) \rightarrow (\text{ReLU}) \rightarrow (1 \times 1 \text{ conv})$ [43]. The attention head is, apart from the number of outputs, structurally equivalent to the semantic head. Furthermore, there is an auxiliary semantic head docking to HRNet, before OCR (not shown in Figure 1).

HMA uses the *region mutual information* (RMI) loss introduced by [50], where \mathcal{L}_{all} composes of a cross-entropy component \mathcal{L}_{ce} and a component representing mutual information (MI) resp. I_l :

$$\mathcal{L}_{\text{all}}(y, p) = \lambda \mathcal{L}_{\text{ce}} + (1 - \lambda) \frac{1}{B} \sum_{b=1}^B \sum_{c=1}^C (-I_l^{b,c}(\mathbf{Y}; \mathbf{P})) \quad (1)$$

$$I_l(\mathbf{Y}; \mathbf{P}) = -\frac{1}{2} \log(\det(\boldsymbol{\Sigma}_{\mathbf{Y}|\mathbf{P}})). \quad (2)$$

λ represents the weighting factor, B the number of batches, C the number of classes. Based on the assumption that pixels do not show local independence, the neighborhood around the pixel is incorporated into the MI computation: \mathbf{Y} and \mathbf{P} form matrices of the ground truth and predictions around the pixel. Equation 2 shows the calculation of RMI by taking the negative log of the determinant of the covariance matrix of \mathbf{Y} and \mathbf{P} . A higher pixel-wise correlation in the region leads to larger determinants, increasing I_l and decreasing the MI component in Equation 1.

During training, data augmentation was used to compensate for the comparatively low amount of data. Scaling, rotation, and shifting are applied to 80% of the samples during training. The images are cropped, if needed, and 20% of the samples obtain 3×3 Gaussian blur. The sampling of patches during training is controlled, such that at 50% of the patches contain defects while the other half does not. Furthermore, boundary tolerance is used in order to account for annotation inaccuracies and uncertainties at class boundaries.

5 Results

In the following, the relevant metrics are introduced, the performance of HMA for different scales is investigated, the attention maps are analyzed, and, finally, benchmarking for crack detection is performed.

5.1 Metrics

Intersection-over-union (aka IoU or Jaccard index) forms the standard evaluation measure for semantic segmentation as applied in benchmarks such as Cityscapes [10] or ADE20K [52]. For evaluating crack segmentation, other measures have been proposed, such as ODS, optimal dataset scale, and OIS, optimal image scale, cf. [54,45,28]. Both metrics compute the F1 score, ODS conditioned by the optimal threshold over the entire dataset and IDS for each individual image being optimally thresholded.

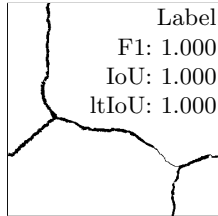
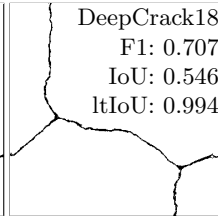
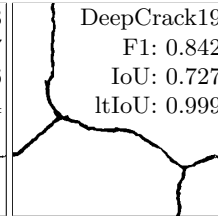
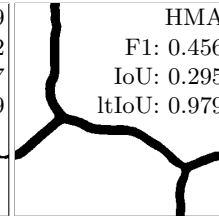
	Label F1: 1.000 IoU: 1.000 ltIoU: 1.000		DeepCrack18 F1: 0.707 IoU: 0.546 ltIoU: 0.994		DeepCrack19 F1: 0.842 IoU: 0.727 ltIoU: 0.999		HMA F1: 0.456 IoU: 0.295 ltIoU: 0.979
---	--	---	--	--	--	---	--

Fig. 2: Evaluation metrics applied on predictions of different approaches. A tolerance of 4 px is used for ltIoU.

Neither IoU nor ODS and OIS appear to be ideal metrics for the assessment of crack detection. Both are sensitive towards the area occupied by a crack, favoring wider compared to more narrow cracks. They consider all pixels, including those at the crack boundary, where class membership is regularly uncertain. As a consequence, an approach is proposed that abstracts from area and takes into account the line-like structure of cracks. For that purpose, the true positives (TP), the false negatives (FN), and the false positives (FP) are computed by:

$$TP = S(T) \cap [S(P) \oplus C(\theta)] \quad (3)$$

$$FP = S(P) \setminus [S(P) \cap [S(T) \oplus C(\theta)]] \quad (4)$$

$$FN = S(T) \setminus [S(T) \cap [S(P) \oplus C(\theta)]] = S(T) \setminus TP \quad (5)$$

T refers to the binary image that represents the ground truth, P to the binary image of predictions, $S(\cdot)$ a skeletonization or thinning method – such as [49,21] – that transforms areas into lines resp. medial axes. Furthermore, C is a circular morphological element with diameter θ that is used for dilation operation \oplus . The diameter θ represents the applied tolerance around the medial axis of each, ground truth and prediction. The *line-based tolerant intersection-over-union* metric is defined as $ltIoU = \frac{TP}{TP+FP+FN}$ using TP, FP, and FN from above.

Figure 2 shows three different predictions for a crack patch, DeepCrack18 [54], DeepCrack19 [29], and the here presented HMA [43]. The top left shows the label. The robustness of ltIoU towards the width of prediction is demonstrated by the relatively stable, close to perfect value of ltIoU, while F1 and IoU vary distinctively. Note that ltIoU is inappropriate if the results from crack segmentation also serve crack width estimation.

5.2 Scales

For assessing the performance of the HMA with respect to all classes, the standard IoU is used. Table 2 shows the IoU for each class conditioned by the scales used for inference. Generally, the inclusion of more scales leads to higher mean IoU, even though the combination of [0.25, 1.0] produces decent results as well.



Fig. 3: Qualitative results of HMA on the test set. Classes: crack (black), spalling (red), corrosion (orange), efflorescence (blue), vegetation (green), and control point (purple).

Table 2: Effects of different scales of HMA on the S2DS dataset.

Dataset		Intersection-over-Union [%]								
		Mean IoU	Background	Crack	Spalling	Corrosion	Efflorescence	Vegetation	Control Point	Runtime (rel.)
Validation	Scales [1.0]	77	96	89	69	79	34	70	99	<i>1.0</i>
	[0.25, 1.0]	86	98	88	80	90	74	72	99	<i>1.3</i>
	[0.5, 1.0]	82	97	90	81	87	53	69	99	<i>1.2</i>
	[0.75, 1.0]	78	96	90	74	85	31	69	99	<i>1.3</i>
	[1.0, 1.5]	75	96	87	64	77	34	70	99	<i>1.6</i>
	[0.5, 1.0, 2.0]	82	97	86	80	87	53	69	99	<i>2.1</i>
	[0.25, 0.5, 1.0]	87	98	87	83	90	76	72	99	<i>1.4</i>
	[0.25, 0.5, 0.75, 1.0]	87	98	87	84	90	76	72	99	<i>1.6</i>
Test	[0.25, 0.5, 1.0]	92	99	91	91	88	90	87	100	–

Seemingly, the scale 0.25 contributes to a better detection of the efflorescence. Unlike for the original HMA – which uses [0.5, 1.0, 2.0] for semantic segmentation of street sceneries – including scale 2.0 does not have a positive impact on performance. An explanation might be that the larger scale does not add information, especially not to the detection of fine structures and boundaries. The overall best combinations are [0.25, 0.5, 1.0] and [0.25, 0.5, 0.75, 1.0]. Due to the slightly lower relative runtime, and the lower memory footprint, [0.25, 0.5, 1.0] is chosen for deployment.

When applied to the test set, Table 2, the overall performance as well as the performance on each class individually improves. This behavior, which is atypical for artificial neural networks, is caused by the higher quality of data in the test set. Due to lack of data, the training and validation set were populated with images of lower quality, in order to hopefully benefit training. Detection on these images was, however, considered optional. The test set, on the other hand, only contains images where detection is considered mandatory. Qualitative results on the test set are presented in Figure 3.

5.3 Attention

Figure 4 illustrates how – mediated by attention – the three different scales contribute to the overall prediction. The top row of each example displays the input image alongside the fused prediction. Below, the attention maps (left) and the corresponding predictions (right) are shown for the three different scales 0.25, 0.5, and 1.0. The attention maps result from a pixelwise softmax across the scales and provide a pixelwise weighting, i.e. the contribution of each pixel of a scale for the overall prediction. Brighter regions in the attention maps refer to higher attention and darker regions to lower attention.

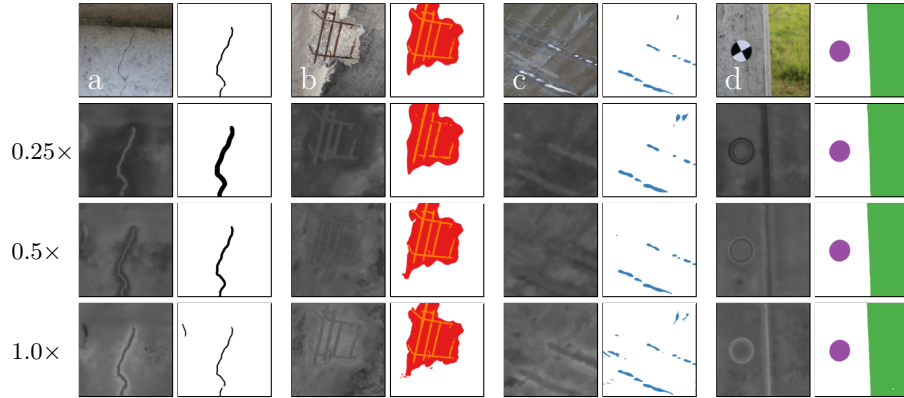


Fig. 4: Attention maps for multiple scales. The input image is displayed in the upper left; right beside the fused prediction. Below the per-scale attention maps are shown with corresponding predictions. Classes: crack (black), spalling (red), corrosion (orange), efflorescence (blue), vegetation (green), and control point (purple).

It can be observed that for all classes, despite spalling, the highest scale ($1.0\times$) shows strong activation of attention around the defect. The defect itself, however, obtains lower attention. On the lowest scale ($0.25\times$) the reverse applies: while the vicinity of the defect shows darker areas which correspond to lower attention, higher attention is paid to the defect itself. This observation applies to crack, corrosion, and efflorescence, whereas control point and vegetation show a similar pattern with less activation on the low scale.

A potential interpretation for this observation is that the highest scales are responsible to determine the point of transition from defect to background for accurate boundary demarcation. Lower scales, on the other hand, possibly rather determine the content of the defect. While the shape of boundary is arbitrary for all defects other than crack and control point, the intensity and color can be characteristic: crack is relatively dark, efflorescence relatively whitish, corrosion has a brownish and vegetation a greenish hue. It is conceivable that these color-related aspects receive more considerations on lower scales.

Contrasting the above observation, spalling seems to show reverse behavior. Lower scales show higher activation in the vicinity and lower activation at the defect itself. On higher scales, the attention at the location of the defect is, however, higher than for other classes. Since the detachment of material causes an edgy and peaky texture in the spalling, it can be conjectured that those relevant details vanish or blur on lower scales. Thus, these features require attention on the highest scale for spalling classification.

5.4 Benchmark

No benchmark is yet available for structural defects as represented in the S2DS dataset. S2DS is rather intended to form a first such benchmark. Certainly, more data need to be acquired and made available to the public, in order to obtain better generalizing models. For crack detection, however, datasets and approaches are available for benchmarking. As pointed out above, IoU does not appear to be a proper metric, which therefore is replaced by ltIoU for the given evaluation.

Table 3: Performance comparison with other approaches and datasets for crack detection. F1 conforms to the ODS metric.

Tol. [px]	Dataset	DeepCrack18		DeepCrack19		HMA	
		F1	ltIoU	F1	ltIoU	F1	ltIoU
0	CRKWH100	0.721	0.564	0.559	0.388	0.521	0.353
	DeepCrack	0.344	0.208	0.881	0.787	0.815	0.688
	S2DS	0.080	0.041	0.347	0.210	0.862	0.758
2	CRKWH100	0.920	0.853	0.735	0.581	0.628	0.458
	DeepCrack	0.405	0.254	0.923	0.857	0.844	0.730
	S2DS	0.156	0.084	0.384	0.238	0.937	0.881
4	CRKWH100	0.951	0.906	0.781	0.641	0.644	0.475
	DeepCrack	0.427	0.271	0.938	0.884	0.856	0.749
	S2DS	0.162	0.088	0.402	0.252	0.949	0.903
8	CRKWH100	0.965	0.932	0.826	0.703	0.659	0.491
	DeepCrack	0.448	0.288	0.952	0.908	0.870	0.769
	S2DS	0.166	0.090	0.430	0.274	0.960	0.922
16	CRKWH100	0.972	0.945	0.873	0.775	0.680	0.516
	DeepCrack	0.474	0.310	0.962	0.928	0.884	0.793
	S2DS	0.170	0.093	0.476	0.312	0.967	0.936
32	CRKWH100	0.976	0.954	0.915	0.844	0.718	0.560
	DeepCrack	0.518	0.349	0.972	0.946	0.901	0.819
	S2DS	0.176	0.096	0.538	0.368	0.972	0.946

Table 3 shows three approaches, DeepCrack18 [54], DeepCrack19 [29], and the here presented HMA [43] applied to three datasets, CRKWH100 [54], DeepCrack [29], and the presented S2DS. The publication of code has not yet become standard practice in crack detection. DeepCrack18 [54] and DeepCrack19 [29], however, are prominent approaches with working code and are used by others for benchmarking, e.g. [28]. CRKWH100 [54] and DeepCrack [29] are the accompanying datasets and regularly serve for benchmarking. The CRKWH100 contains thin pavement cracks, DeepCrack covers various types of cracks, and S2DS mainly represents cracks in concrete walls. Six levels of tolerance for the positioning of the medial axis are investigated.

Generally, F1 and ltIoU improve with higher tolerance. Even though at tolerance level 16 and 32 px saturation effects can be observed, i.e. more tolerance

does not lead to distinctively better performance. This point can be considered the currently best possible performance of the classifier. All approaches perform best on the datasets created in their context. There are, however, differences with respect to the generalization capabilities. While DeepCrack18 shows at most mediocre performance on datasets other than CRKWH100, the performance of HMA also deteriorates on the other data, but less severely. On CRKWH100, the HMA regularly shows activation for the spalling class. This confusion might be rooted in the occasional textural similarity of pavement and spalling. It, however, raises the question, if pavement crack detection and spalling detection can be reasonably represented in a single approach. Based on the benchmarking results can be stated that a domain gap exists, particularly with respect to pavement cracks, which, though, might be bridgeable even with HMA.

6 Conclusion

In the context of this work, a dataset was created for multi-class classification of several structural defects. This dataset is unprecedented in quantity and quality and is intended to form a first benchmarking dataset in the domain. While other researchers focus on the design of suitable artificial neural network architectures, it is demonstrated that by means of transfer learning on the state-of-the-art approach of hierarchical multi-scale attention [43] a decent performance can be achieved. Network design can, however, be appropriate if a relatively small, application-tailored model is required, e.g. for deployment in an embedded platform on a UAS.

Furthermore, the investigation of attention revealed the relevance of large scales for demarcating class boundaries. Smaller scales, on the other hand, show higher activation directly at the defect, which led to the conjecture, that they contribute to determining the proper class. Justifications for spalling not conforming to this pattern was given by the high degree of details in the textural pattern only perceivable on larger scales.

It is claimed, that standard evaluation metrics, such as F1 and IoU, are not appropriate for evaluating crack segmentation. The measures conceptually evaluate the overlap of areas. Cracks are, however, rather line-like structures and, thus, require other metrics for plausible comparison. For that purpose, the ltIoU was introduced, which reduces a prediction to a medial axis and assesses the intersection and union of medial axes of ground truth and prediction given a certain positional tolerance. Note that the measure is, however, unsuited if the predictions are directly used for crack width estimation. By means of ltIoU an intra-domain gap could be observed in benchmarking: the performance of crack detection very much depends on the data available during training. This holds for the background, e.g. pavement being mistaken for spalling, or distractive artifacts, such as concrete texture falsely classified as cracks. To come up with a general approach to crack detection covering various different surfaces and crack types remains an open challenge.

Acknowledgment

The authors would like to thank *DB Netz AG* and *Leonhardt, Andrä und Partner (LAP)* for providing numerous images as well as their consent to publication. Without them this work would have been impossible.

References

1. Atha, D.J., Jahanshahi, M.R.: Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection. *Structural Health Monitoring* **17**(5), 1110–1128 (2018)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence* **39**(12), 2481–2495 (2017)
3. Benz, C., Rodehorst, V.: Model-based crack width estimation using rectangle transform. In: 17th International Conference on Machine Vision and Applications (MVA). pp. 1–5. IEEE (2021)
4. Bianchi, E., Abbott, A.L., Tokekar, P., Hebdon, M.: Cocombridge: Structural detail data set for bridge inspections. *Journal of Computing in Civil Engineering* **35**(3), 04021003 (2021)
5. Borse, S., Wang, Y., Zhang, Y., Porikli, F.: Inverseform: A loss function for structured boundary-aware segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5901–5911 (2021)
6. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)
7. Cha, Y.J., Choi, W., Suh, G., Mahmoudkhani, S., Büyüköztürk, O.: Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering* **33**(9), 731–747 (2018)
8. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(4), 834–848 (2017)
9. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017)
10. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3213–3223 (2016)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. IEEE (2009)
12. Dorafshan, S., Thomas, R.J., Maguire, M.: Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete. *Construction and Building Materials* **186**, 1031–1045 (2018)
13. Dorafshan, S., Thomas, R.J., Maguire, M.: Sdnet2018: An annotated image dataset for non-contact concrete crack detection using deep convolutional neural networks. *Data in brief* **21**, 1664–1668 (2018)

14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
15. Dung, C.V., et al.: Autonomous concrete crack detection using deep fully convolutional neural network. *Automation in Construction* **99**, 52–58 (2019)
16. Duy, L.D., Anh, N.T., Son, N.T., Tung, N.V., Duong, N.B., Khan, M.H.R.: Deep learning in semantic segmentation of rust in images. In: *Proceedings of the 9th International Conference on Software and Computer Applications*. pp. 129–132 (2020)
17. Forkan, A.R.M., Kang, Y.B., Jayaraman, P.P., Liao, K., Kaul, R., Morgan, G., Ranjan, R., Sinha, S.: Corrdetector: A framework for structural corrosion detection from drone images using ensemble deep learning. arXiv preprint arXiv:2102.04686 (2021)
18. Gao, Y., Mosalam, K.M.: Deep transfer learning for image-based structural damage recognition. *Computer-Aided Civil and Infrastructure Engineering* **33**(9), 748–768 (2018)
19. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3354–3361. IEEE (2012)
20. Goodfellow, I., Bengio, Y., Courville, A.: *Deep learning*. MIT press (2016)
21. Guo, Z., Hall, R.W.: Parallel thinning with two-subiteration algorithms. *Communications of the ACM* **32**(3), 359–373 (1989)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 770–778 (2016)
23. Hoskere, V., Narazaki, Y., Hoang, T., Spencer Jr, B.: Vision-based structural inspection using multiscale deep convolutional neural networks. arXiv preprint arXiv:1805.01055 (2018)
24. Hoskere, V., Narazaki, Y., Hoang, T.A., Spencer Jr, B.: Madnet: multi-task semantic segmentation of multiple types of structural materials and damage in images of civil infrastructure. *Journal of Civil Structural Health Monitoring* **10**, 757–773 (2020)
25. Katsamenis, I., Protopapadakis, E., Doulamis, A., Doulamis, N., Voulodimos, A.: Pixel-level corrosion detection on metal constructions by fusion of deep learning semantic and contour segmentation. In: *International Symposium on Visual Computing*. pp. 160–169. Springer (2020)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* **25**, 1097–1105 (2012)
27. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *European Conference on Computer Vision*. pp. 740–755. Springer (2014)
28. Liu, H., Miao, X., Mertz, C., Xu, C., Kong, H.: Crackformer: Transformer network for fine-grained crack detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3783–3792 (2021)
29. Liu, Y., Yao, J., Lu, X., Xie, R., Li, L.: Deepcrack: A deep hierarchical feature learning architecture for crack segmentation. *Neurocomputing* **338**, 139–153 (2019)
30. Liu, Z., Cao, Y., Wang, Y., Wang, W.: Computer vision-based concrete crack detection using u-net fully convolutional networks. *Automation in Construction* **104**, 129–139 (2019)

31. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and pattern recognition. pp. 3431–3440 (2015)
32. Mohan, A., Poobal, S.: Crack detection using image processing: A critical review and analysis. *Alexandria Engineering Journal* **57**(2), 787–798 (2018)
33. Narazaki, Y., Hoskere, V., Yoshida, K., Spencer, B.F., Fujino, Y.: Synthetic environments for vision-based structural condition assessment of japanese high-speed railway viaducts. *Mechanical Systems and Signal Processing* **160**, 107850 (2021)
34. Ortiz, A., Bonnin-Pascual, F., Garcia-Fidalgo, E., et al.: Vision-based corrosion detection assisted by a micro-aerial vehicle in a vessel inspection application. *Sensors* **16**(12), 2118 (2016)
35. Ortiz, A., Bonnin-Pascual, F., Garcia-Fidalgo, E., et al.: Visual inspection of vessels by means of a micro-aerial vehicle: an artificial neural network approach for corrosion detection. In: Robot 2015: Second Iberian Robotics Conference. pp. 223–234. Springer (2016)
36. Pan, X., Yang, T.: Postdisaster image-based damage detection and repair cost estimation of reinforced concrete buildings using dual convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering* **35**(5), 495–510 (2020)
37. Pauly, L., Hogg, D., Fuentes, R., Peel, H.: Deeper networks for pavement crack detection. In: Proceedings of the 34th ISARC. pp. 479–485. IAARC (2017)
38. Perez, H., Tah, J.H., Mosavi, A.: Deep learning for detecting building defects using convolutional neural networks. *Sensors* **19**(16), 3556 (2019)
39. Petricca, L., Moss, T., Figueroa, G., Broen, S.: Corrosion detection using ai: a comparison of standard computer vision techniques and deep learning model. In: Proceedings of the Sixth International Conference on Computer Science, Engineering and Information Technology. vol. 91, p. 99 (2016)
40. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention. pp. 234–241. Springer (2015)
41. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
42. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and pattern recognition. pp. 1–9 (2015)
43. Tao, A., Sapra, K., Catanzaro, B.: Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821 (2020)
44. Xu, Y., Xiao, T., Zhang, J., Yang, K., Zhang, Z.: Scale-invariant convolutional neural networks. arXiv preprint arXiv:1411.6369 (2014)
45. Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X., Ling, H.: Feature pyramid and hierarchical boosting network for pavement crack detection. *IEEE Transactions on Intelligent Transportation Systems* **21**(4), 1525–1535 (2019)
46. Yang, X., Li, H., Yu, Y., Luo, X., Huang, T., Yang, X.: Automatic pixel-level crack detection and measurement using fully convolutional network. *Computer-Aided Civil and Infrastructure Engineering* **33**(12), 1090–1109 (2018)
47. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: ECCV 2020: 16th European Conference on Computer Vision, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. pp. 173–190. Springer (2020)

48. Zhang, L., Yang, F., Zhang, Y.D., Zhu, Y.J.: Road crack detection using deep convolutional neural network. In: IEEE International Conference on Image Processing (ICIP). pp. 3708–3712. IEEE (2016)
49. Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. *Communications of the ACM* **27**(3), 236–239 (1984)
50. Zhao, S., Wang, Y., Yang, Z., Cai, D.: Region mutual information loss for semantic segmentation. arXiv preprint arXiv:1910.12037 (2019)
51. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6881–6890 (2021)
52. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision* **127**(3), 302–321 (2019)
53. Zou, Q., Cao, Y., Li, Q., Mao, Q., Wang, S.: Cracktree: Automatic crack detection from pavement images. *Pattern Recognition Letters* **33**(3), 227–238 (2012)
54. Zou, Q., Zhang, Z., Li, Q., Qi, X., Wang, Q., Wang, S.: Deepcrack: Learning hierarchical convolutional features for crack detection. *IEEE Transactions on Image Processing* **28**(3), 1498–1512 (2018)