dacl-challenge: Semantic Segmentation during Visual Bridge Inspections

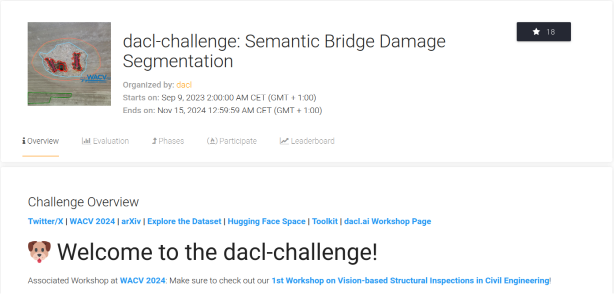
Johannes Flotzinger[1], Philipp J. Rösch[1], Christian Benz[2],
Muneer Ahmad[3], Murat Cankaya[1], Helmut Mayer[1],
Volker Rodehorst[2], Norbert Oswald[1], Thomas Braml[1]

[1]University of the Bundeswehr Munich, Germany, {firstname.lastname}@unibw.de
[2]Bauhaus-Universität Weimar, Germany, {firstname.lastname}@uni-weimar.de
[3]NetApp, Germany, muneer.ahmad@netapp.com

Abstract

Civil engineering structures – such as bridges – form an essential component of the transportation infrastructure. A failure of an individual structure can result in enormous damage and costs. The economic costs caused by the closure of a bridge due to congestion can be many times the costs of the bridge itself and its maintenance. Thus, it is mandatory to keep these structures in a safe and operational state. In order to ensure this, they are frequently inspected. However, the current inspection process is error-prone and lengthy. Especially the damage documentation using a hand-drawn sketch causes inconsistencies in the building assessment. On the other hand, recent advancements in hardware enable the deployment of computer vision models for increasing the quality, traceability, and efficiency of structural inspections. Such models are the key element of digitized structural inspections and the basis for automated damage classification, measurement and localization on a pixel-level. Current datasets available for this task suffer from limitations in both size and diversity of classes, raising concerns about their applicability in real-world contexts and their effectiveness as benchmarks. Addressing this problem, we introduced "dacl10k" (damage classification), a diverse dataset designed for multi-label semantic segmentation. Comprising 9,920 images extracted from real-world bridge inspections, "dacl10k" stands out by its comprehensive coverage. It includes 13 damage classes and 6 crucial bridge components pivotal in assessing structures and guiding decisions on restoration, traffic restrictions, and bridge closures. To accelerate progress in baseline development, we organized the "dacl-challenge", inviting enthusiasts in damage recognition to vie for training the best performing model on the "dacl10k" dataset. The competition is at the core of the "1st Workshop on Vision-Based Structural Inspections in Civil Engineering", hosted at WACV 2024.

In total, 23 participants registered for the challenge, with eight achieving a performance superior to our baseline. The best result shows a mean intersection-over-union of 51%. This paper delineates the challenge's structure, introduces the dataset utilized, presents the achieved outcomes, and outlines prospective avenues for further exploration in this domain.



dacl-challenge: Semantic Bridge Damage Segmentation

Organized by: dacl
Starts on: Sep 9, 2023 2:00:00 AM CET (GMT + 1:00)
Ends on: Nov 15, 2024 12:59:59 AM CET (GMT + 1:00)

Challenge Overview

Twitter/X | WACV 2024 | arXiv | Explore the Dataset | Hugging Face Space | Toolkit | dacl.ai Workshop Page

🐶 Welcome to the dacl-challenge!

Associated Workshop at WACV 2024: Make sure to check out our 1st Workshop on Vision-based Structural Inspections in Civil Engineering!

Figure 1. Screenshot of the eval.ai challenge website.

1. Introduction

Civil engineering structures are the key element of public infrastructure. They can be subdivided into the fields of transportation (e.g., road and railway bridges, tunnels, airports, harbors), energy supply (for instance, wind turbines, offshore constructions, power plants), water supply (for example dams, pumping stations, pipelines) and waste management (*e.g*., sewers, wastewater and storage facilities). It is of great importance for our daily life that these structures operate technically faultlessly and their maintenance measures are well planned. A failure of an individual structure can result in enormous damage and costs. For example, the economic damage caused by the closure of a bridge due to congestion can be many times the cost of the bridge itself and its maintenance. To avoid this, such structures need to

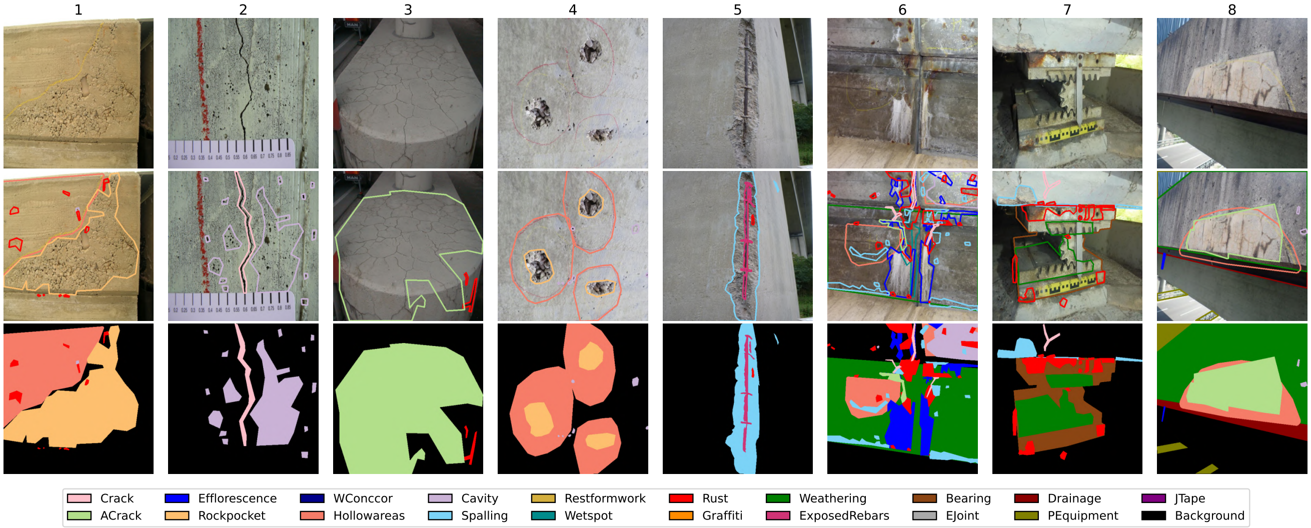


Figure 2. Example annotations from dacl10k. Top row: original image. Middle row: polygonal annotations. Bottom row: stacked masks. From left to right, the images display the individual classes: 1. *Rockpocket, Hollowarea, Rust, Cavities*; 2. *Crack, Cavities*; 3. *Rust, Alligator Crack*; 4. *Hollowareas, Rockpockets, Cavities*; 5. *Spalling, Exposed Rebars, Rust*; 6. *Efflorescence, Rust, Spallings, Cracks, Wetspot, Cavities, Hollowarea, Weathering, ACrack*; 7. *Rust, Weathering, Bearing, Crack, Spallings*; 8. *Protective Equipment, Efflorescence, Drainage, Cavities, Hollowarea, Weathering, ACrack*. The following classes are abbreviated: *Alligator Crack* (ACrack), *Washouts/Concrete corrosion* (WConccor), *Expansion Joint* (EJoint), *Protective Equipment* (PEquipment) and *Joint Tape* (JTape).

be frequently inspected, in order to: (*i*) plan maintenance works, (*ii*) prevent abrupt failure, and (*iii*) avoid unnecessary closures or load limitations.

Facing a continually growing percentage of infrastructure reaching a critical age concerning the occurrence of damage, their inspection is more important than ever. In the case of bridges, many governmental reports describe the current state as severe. For instance, “more than 222,000 U.S. bridges need major repair work or should be replaced [In 2023]” [1] or “... at least 25,000 road bridges [in France] are in poor structural condition ...” [4]. At the same time, governments have to deal with a lack of funding and the construction industry is confronted with a reduced availability of skilled staff [3]. This underlines the demand for a more efficient examination pipeline.

However, the current inspection process for civil engineering structures, consisting of damage recognition, assessment and documentation, is mostly performed in an analogue way. In the case of bridge inspection, this means that an inspector visually recognizes each defect, noting the damage class, values of measurements, and the location in a 2D sketch. In addition, the inspector assigns an image filename in order to combine the defect information with the corresponding image later when writing the assessment, or rather the inspection report.

Using computer vision methods to automatically classify, measure and localize defects and building components as part of smartphone applications or software for unmanned aircraft systems (UAS) makes the inspection of such infrastructure much more efficient [27]. As damage recognition and documentation is in large parts performed automatically by software, the inspector can concentrate on damage assessment and interactions between defects. In addition, the assessment can be done more frequently. This leads to a higher confidence in the determination of structural integrity, traffic safety and long-term durability. Additionally, a higher frequency of inspections allows for extending the useful life of buildings in a critical condition, or avoiding/postponing load limitations (e.g., ban of heavy traffic on bridges). Furthermore, capturing the damage (and objects) may be executed by non-professionally trained personnel, which can compensate for the problem of staff shortages to some extent. This problem is present in many countries such as Germany, France, and the US.

Compared to other domains, such as autonomous driving [7, 11, 15, 30, 35] or medicine [26, 29, 31], infrastructure defect recognition received only little attention in the past. Furthermore, the existing datasets for this particular task are constrained by their limited size and the lack of diversity in the classes. This raises doubts about how well they comply with real-world situations and questions their efficacy as benchmarks.

A central contribution to this problem is the dacl10k dataset [14], which includes images collected during inspections of concrete bridges acquired from databases at authorities and engineering offices, thus, representing real-world scenarios. Concrete bridges are the most common bridge type along with steel, steel composite, and wooden bridges. The dataset provides polygonal annotations for 13 bridge defects as well as six bridge components that play a key role in the structure assessment. Two groups of annotators, civil engineering students and a professional annotation team, tagged the images with a semantic segmentation by hand. Yet, dacl10k is not restricted to concrete bridges. Its concrete and general defect groups can appear on any building made of concrete (*e.g.*, *Crack*, *Spalling*, etc.), and some on steel structures (for instance, *Rust*, *Graffiti*, etc.). Therefore, it is relevant for most civil engineering structures.

To popularize dacl10k, we organize the "dacl-challenge", which aims to find the best multi-label semantic segmentation models for the novel, highly diverse, large-scale dataset. The aim of the challenge is to provide a benchmarking platform for the automatic visual inspections of bridges. The platform of the challenge will be maintained also after completion of the challenge for future benchmarking.

The challenge is the central part of the "1st Workshop on Vision-Based Structural Inspections in Civil Engineering". It focuses on the visual recognition of defects and building components utilizing innovative computer vision methods to increase the efficiency of the laborious inspection process of civil engineering structures. We invite experts who successfully employ computer vision for visual inspections to present their applications and emerging challenges at the workshop.

The challenge and the workshop will highlight the yet mostly unnoticed problem of visual structural inspections for public infrastructure. Furthermore, a community for the field of computer vision-based inspections of civil engineering structures is to be created promoting research in this field. The gain in efficiency for inspections will lead to safer and more cost-effective public infrastructure.

2. Related Work

Damage Recognition on Built Structures. In recent years, the application of damage recognition on built structure has evolved. Former datasets and models focus on image classification only. Several datasets deal with binary classification (mostly crack vs. no crack) [19, 23] and a few also with multiple classes [18, 28]. Recently, the focus has moved towards pixel-level segmentation tasks, which is more challenging but also more helpful for the inspection process. Important work in this field is given by CrackSeg9k [22] and UAV75 [5] for semantic crack segmentation and S2DS [6] for multi-class semantic segmentation.

Multi-label Semantic Segmentation. Semantic Segmentation is the task to predict object classes in images on pixel level. Many powerful models have been developed in recent years, e.g. DeepLabV3+ [8, 9], Feature Pyramid Network (FPN) [20] or SegFormer [34]. But with most datasets, there is only one correct class per pixel. In our case, a pixel can belong to several classes. For example, there may be *Spalling* that shows also signs of *Rust*. Therefore, we are no longer referring to multi-class semantic segmentation, but to multi-label semantic segmentation.

Workshops. Civil engineering, particularly visual inspection has recently increased its visibility in the form of workshops at computer vision conferences. For example, at ECCV2022 the *Computer Vision for Civil and Infrastructure Engineering Workshop* [16] was organized. The workshop is motivated by "Civil and infrastructure engineering are corner stones in modern society". It aims at different built structures such as bridges, roads, sewerage, and buildings. Computer vision is seen as an important component for automated inspection, but it is also emphasized that it can "analyze work patterns or detect hazardous situations at, e.g., construction sites".

The *1st workshop on Vision-based InduStrial InspectiON (VISION)* [2] strived to be a platform for sharing improvements in research and addressing emerging practical challenges in industrial inspection using computer vision. It included the two challenges object detection for industrial products (metal cylinder, rings, etc.) and object generation.

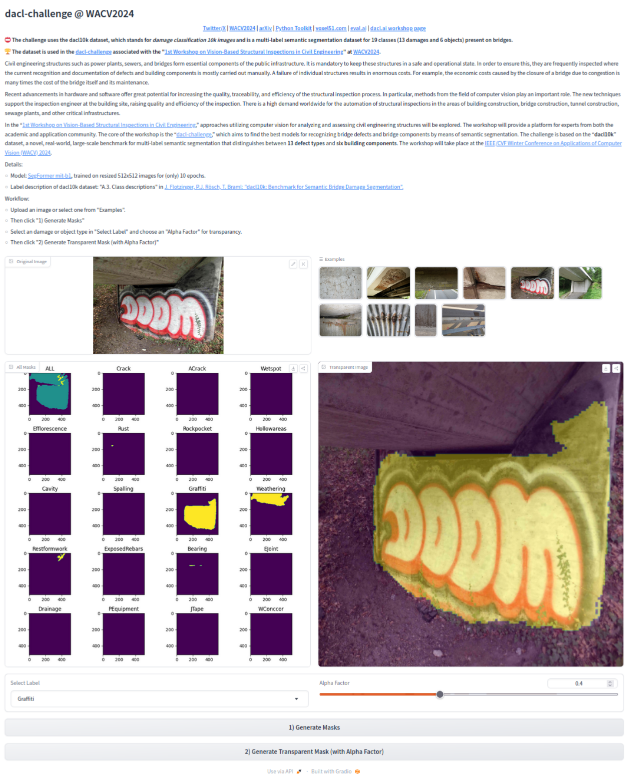
3. Challenge Design

After our challenge was accepted on August 22, 2023, we set it up on eval.ai[1] and started on September 9, 2023. See Fig. 1 for a screenshot of the challenge website and Tab. 1 for a detailed timetable. In Tab. 2 you can see which release the splits belong to and which splits were used for the leaderboard.

For promotion a Hugging Face Space[2] was created to give an impression how results can look like (see Fig. 3). Moreover, an interactive data visualization tool was made available via Voxel51.[3] The team also provided a Python toolkit[4] to lower the entry barrier to the challenge. Participants were kept up to date via X.[5] The challenge was split into two phases: the development phase and the final test phase.

Development Phase. In the development phase, images for the training (n=6,935), validation (n=975) and testing

[1] https://eval.ai/web/challenges/challenge-page/2130/overview
[2] https://huggingface.co/spaces/phiyodr/dacl-challenge
[3] https://try.fiftyone.ai/datasets/dacl10k/samples
[4] https://github.com/phiyodr/dacl10k-toolkit
[5] https://x.com/dacl_ai/



dacl-challenge @ WACV2024

Twitter/X | WACV2024 | arXiv | Python Toolkit | voxel51.com | eval.ai | dacl.ai workshop page

The challenge uses the dacl10k dataset, which stands for *damage classification 10k images* and is a multi-label semantic segmentation dataset for 19 classes (13 damages and 6 objects) present on bridges.

The dataset is used in the dacl-challenge associated with the "1st Workshop on Vision-Based Structural Inspections in Civil Engineering" at WACV2024.

Civil engineering structures such as power plants, sewers, and bridges form essential components of the public infrastructure. It is mandatory to keep these structures in a safe and operational state. In order to ensure this, they are frequently inspected where the current recognition and documentation of defects and building components is mostly carried out manually. A failure of individual structures results in enormous costs. For example, the economic costs caused by the closure of a bridge due to congestion is many times the cost of the bridge itself and its maintenance.

Recent advancements in hardware and software offer great potential for increasing the quality, traceability, and efficiency of the structural inspection process. In particular, methods from the field of computer vision play an important role. The new techniques support the inspection engineer at the building site, raising quality and efficiency of the inspection. There is a high demand worldwide for the automation of structural inspections in the areas of building construction, bridge construction, tunnel construction, sewage plants, and other critical infrastructures.

In the "1st Workshop on Vision-Based Structural Inspections in Civil Engineering," approaches utilizing computer vision for analyzing and assessing civil engineering structures will be explored. The workshop will provide a platform for experts from both the academic and application community. The core of the workshop is the "dacl-challenge," which aims to find the best models for recognizing bridge defects and bridge components by means of semantic segmentation. The challenge is based on the "**dacl10k**" dataset, a novel, real-world, large-scale benchmark for multi-label semantic segmentation that distinguishes between **13 defect types** and **six building components**. The workshop will take place at the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) 2024.

Details:

- Model: SegFormer mit-b1, trained on resized 512x512 images for (only) 10 epochs.
- Label description of dacl10k dataset: "A.3. Class descriptions" in J. Flotzinger, P.J. Rösch, T. Braml: "dacl10k: Benchmark for Semantic Bridge Damage Segmentation".

Workflow:

- Upload an image or select one from "Examples".
- Then click "1) Generate Masks"
- Select an damage or object type in "Select Label" and choose an "Alpha Factor" for transparancy.
- Then click "2) Generate Transparent Mask (with Alpha Factor)"

Select Label: Graffiti

Alpha Factor: 0.4

Figure 3. Hugging Face Space with dacl-challenge demo using SegFormer MiT-b1.

(n=1,012) split were made available via AWS and GigaMove. For the training and validation split, additionally annotation files in a labelme-like format[6] were provided. Participants could upload their results to eval.ai to receive feedback about their model performance. Some results were publicly visible in the leaderboard, others opted for a private listing only. We also provided a baseline (cf. Sec. 3.3) to act as a reference point for the participants. The development phase was active until October 27, 2023 (48 days).

Testfinal Phase. In the second phase another 998 images – called "testchallenge" – were provided. The task in this phase was to create predictions for the images from testdev and testchallenge combined, termed "Testfinal". Participants could only see their own results and could no longer compare themselves with others. Due to server problems at eval.ai, the end of the challenge had to be postponed from November 10th to November 14th. This phase has been active for 18 days.

September 9, 2023	Start of Development Phase
October 27, 2023	Start of Final Phase
November 14, 2023	End of Challenge
November 14, 2023	Fact-Sheet Submission
November 21, 2023	Release of Final Results
January 7, 2024	Workshop and Prize Ceremony

Table 1. Timetable of the dacl-challenge.

	Released in phase		Used in leaderboard	
Split	Devel.	Testfinal	Devel.	Testfinal
train	✓			
validation	✓			
testdev	✓		✓	✓
testchallenge		✓		✓

Table 2. Splits released and used in different phases.

Fact Sheet. In order to win prizes, a fact sheet[7] had to be submitted via the Uni-Weimar cloud. Participants had to explain their approach on two pages and provide a variety of relevant parameters concerning their best model. We also encouraged everyone to openly share their code.

Workshop. The workshop is scheduled to take place on January 7, 2024. It will include keynote presentations and the prize ceremony at which the three best results will be honored. Prize money of $3,000, $2,000 and $1,500, respectively, will be awarded. To this end, the ranking on the "Testfinal" leaderboard for the mIoU metric (see Sec. 3.2) is used combined with the submission of a proper fact sheet. After the end of the challenge, we have moved the evaluation process to CodaLab[8] so that the testdev split can still be used in research.

3.1. Challenge Dataset: dacl10k-v2

The challenge dataset is the second version of dacl10k, a multi-label segmentation dataset consisting of (mostly smartphone) images from bridge inspections in Germany (see Figure 2). Its images are highly diverse with respect to the lighting condition, camera pose and resolution. The annotations were created by civil engineering students and an external team using a detailed annotation guideline.

"dacl10k-v2" is an improved version of the initial dataset, which was introduced in Flotzinger *et al.* [14]. Apart from a general cleaning, the main difference between v2 and v1 is the separation of the *Rockpocket* and *Cavity* defect class. In Figure 2 example 1 shows *Rockpocket* in pale orange and example 2 displays *Cavity* in pale purple. This step has been taken due to the fact that in [14] weak performance has been reported on the *Rockpocket* damage (29%), noting that from the application point of view a further differentiation would be beneficial. In the previous version,

[6] https://github.com/wkentaro/labelme

[7] https://dacl.ai/assets/call-for-fact-sheets.pdf

[8] https://codalab.lisn.upsaclay.fr/competitions/16317

both defects were included in *Rockpocket*. They were originally merged in one class as their cause is similar, namely insufficient deaeration of the concrete. Additionally, a *Cavity* looks, when zoomed in, like a small *Rockpocket*. So, they mainly differ in size.

Figure 4 shows the significantly differing number of pixels for each class in the dataset. The classes *Weathering* and *PEquipment* have the highest number of pixels under all the classes within the dacl10k dataset, while *Exposed Rebars*, *Joint Tape*, *Rockpocket* and *Restformwork* show the lowest number of pixels.

Tab. 3 depicts a more detailed insight about the dataset. While *Cavity* has the topmost number of instances, *ACrack* has only a few instances. Yet, *ACrack* shows the highest pixels to polygon and pixel to image ratio – hence indicate large surface damages. The damage with the smallest area per polygon is *Cavity*. Furthermore, *Crack* is the damage with the lowest number of pixels per image. Regarding the major change from dacl10k-v1 to dacl10k-v2 (splitting the *Rockpocket* class) it can be noted that the share of this class with respect to the total pixel area was reduced from 1.48% to 0.14%. The share of *Cavity* in the second version amounts 0.39% signifying that many false positives were deleted.

The data files for the development phase[9] and the final test phase[10] are available online.

3.2. Evaluation Protocol

In the context of the challenge the standard metric for semantic segmentation is used, the *Intersection-over-Union* (IoU, also referred to as *Jaccard index*). Since multiple labels for each pixel are allowed, the IoU is computed class-wise and subsequently aggregated to the *mean IoU* (mIoU). The mIoU is the (unweighted) arithmetic average of the class IoUs. The metrics are defined as:

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N} \text{IoU}_i \tag{2}$$

TP, FP, and FN refer to the true positives, false positives, and false negatives, respectively. N represents the number of classes, here $N = 19$, and IoU_i refers to the IoU of the i-th class. The ranking and challenge winner is determined based on the highest mIoU.

3.3. Baseline

In [14], three architectures were compared. These were DeepLabV3+ and FPN with MobileNetV3-Large, EfficientNet-B2 and EfficientNet-B4 encoders respectively, and SegFormer MiT-b1. Each model was trained on dacl10k-v1 dataset with different learning rates and only the best results were reported.

For the challenge we use “dacl10k-v2” (cf. above). We

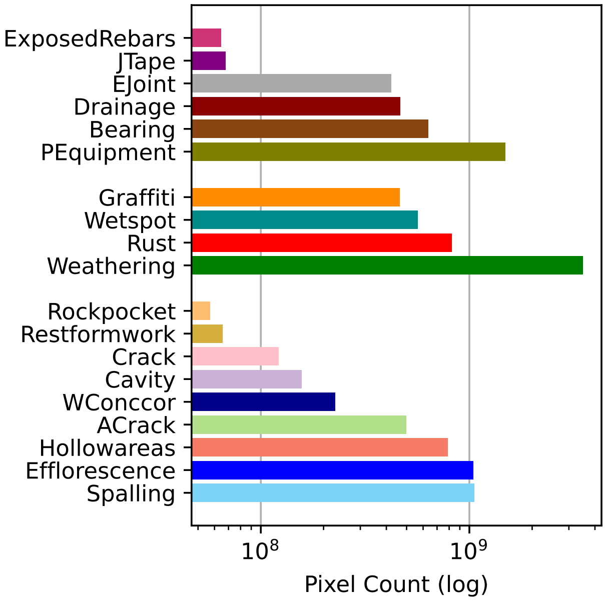


Figure 4. Pixel counts with respect to each class in dacl10k based on the original image sizes. The bars are arranged according to group affiliations.

Class	#polyg./ image	#pixels/ polyg.	#pixels/ image	%polyg.	%pixels
Crack	1.8	27,434	49,503	4.07	0.3
ACrack	1.12	947,430	1,062,822	0.48	1.23
Efflorescence	2.3	208,256	478,408	4.6	2.59
Rockpocket	1.74	71,039	123,780	0.74	0.14
WConccor	1.35	457,780	617,008	0.45	0.56
Hollowareas	1.21	415,309	504,532	1.74	1.95
Cavity	6.77	13,470	91,182	10.67	0.39
Spalling	2.61	85,484	223,106	11.33	2.62
Restformw.	1.2	50,146	59,929	1.2	0.16
Wetspot	1.47	273,825	403,511	1.89	1.4
Rust	3.62	46,640	168,782	16.2	2.04
Graffiti	2.29	172,849	396,007	2.47	1.15
Weathering	1.43	615,624	881,279	5.21	8.66
ExposedR.	2.25	25,770	58,035	2.29	0.16
Bearing	1.45	422,297	613,103	1.38	1.57
EJoint	1.12	701,120	784,370	0.55	1.05
Drainage	1.37	230,665	316,051	1.85	1.15
PEquipment	1.22	641,596	785,870	2.12	3.67
JTape	1.19	49,370	58,529	1.26	0.17
Background	3.3	925,898	3,051,636	29.49	73.72

Table 3. Overall statistics of the dataset regarding average number of polygons per image, number of pixels per polygon, number of pixels per image, share of polygons and share of pixels. Midrules separate the classes according to their group affiliations.

[9] https://dacl10k.s3.eu-central-1.amazonaws.com/dacl10k-challenge/dacl10k_v2_devphase.zip (4.8 GB)

[10] https://dacl10k.s3.eu-central-1.amazonaws.com/dacl10k-challenge/dacl10k_v2_testchallenge.zip (0.5 GB)

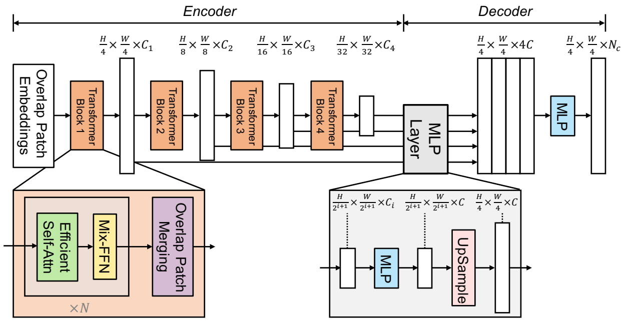

Encoder
Decoder
$\frac{H}{4}\times\frac{W}{4}\times C_1$
$\frac{H}{8}\times\frac{W}{8}\times C_2$
$\frac{H}{16}\times\frac{W}{16}\times C_3$
$\frac{H}{32}\times\frac{W}{32}\times C_4$
$\frac{H}{4}\times\frac{W}{4}\times 4C$
$\frac{H}{4}\times\frac{W}{4}\times N_c$
Overlap Patch Embeddings
Transformer Block 1
Transformer Block 2
Transformer Block 3
Transformer Block 4
MLP Layer
MLP
Efficient Self-Attn
Mix-FFN
Overlap Patch Merging
$\times N$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i$
$\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C$
$\frac{H}{4}\times\frac{W}{4}\times C$
MLP
UpSample

Figure 5. SegFormer architecture. Picture credit: [34].

only provide one model, which serves as a simple baseline and reference point for the participants.

As baseline SegFormer MiT-b1 [34] is used, which has a relatively small 13.1 million parameter encoder with a mulit-label segmentation head. The model was pre-trained on the ImageNet-1k dataset and the architecture can be investigated in Fig. 5. For the baseline, SegFormer was trained for 10 epochs for the development leaderboard (mIoU of 34.6%) and 30 epochs for the Testfinal leaderboard (37.2%). All other training parameters remained unchanged from [14]. In Tab. 4 we named the baseline *dacl-squad*.

4. Challenge Results

In this section the results of the challenge are presented. Furthermore, information about the submitted approaches is provided.

4.1. Leaderboard

Testfinal. Tab. 4 shows the state of the leaderboard at the end of the challenge. The winner is determined based on the Testfinal leaderboard. With an mIoU of 51% *Sheoran* wins the dacl-challenge at WACV 2024 followed by *Bridge Protector* (50.5%) and *Winning Wieners* (50%). *Sheoran* performs on top of five of the 19 classes. The runner-up *Bridge Protector* performs best on six classes, however, showing comparatively low performance on the Washouts/Concrete corrosion and Wetspot class. It performs noticeably well on the Crack class surpassing all others by 4.5% points. *Sheoran* shows a relatively balanced performance over all classes and performs particularly well on the Wetspot and Drainage class, which paved the road to victory. The mIoU of the top 3 is within a range of 1% indicating that nuances lead to the final ranking.

In terms of class performance, Washouts/Concrete corrosion is found to be particularly difficult. As pointed out by *Bridge Protector*, this class is underrepresented in the training set. Furthermore, Cavity and Crack are challenging classes for all approaches. PEquipment is the class best predicted across all approaches peaking at 83.1% for *Bridge Protector*.

Development. The bottom part of Tab. 4 provides insights into the performance of the participants during the development phase. Some participants including *Bridge Protector* and *mp269546* did not submit results in the development phase. Apart from *Shivesh Khaitan*, the order of the other approaches corresponds to the Testfinal set, with *Sheoran* also ranking first.

4.2. Overview

Despite noticeable overlap in the architectures, the landscape of the top 10 approaches is diverse with respect to the choice of backbones and training configurations. Many submissions are based on established segmentation models such as FPN, SegFormer, ConvNeXt, UPerNet, U-Net, and Mask2Former. Others also explore EfficientNet and Yolo-v8. Winning Wieners propose a novel method by combining FPN with MaxViT. Used backbones include EVA-02, ConvNeXt, InternImage, MaxViT, and BEiT, mostly pre-trained on ADE20K or ImageNet. The submissions usually rely on standard augmentation strategies including crops and flips. *Sheoran* emphasizes the use of CutMix and Winning Wieners apply RandAugment. Color jittering showed minor effectiveness especially for color sensitive classes, as mentioned by *mp269546*.

A recurring and effective feature is the use of ensembles. *Sheoran* selectively groups together predictions from multiple models while *Winning Wieners* combine the predictions trained on different folds of the dataset. A number of approaches apply test time augmentation (TTA). The used losses span from BCE, Jaccard, Dice over recall, focal up to Lovasz loss. AdamW is the prevailing optimizer, only *Sheoran* (RangerLars) and *Winning Wieners* (MADGRAD [12]) use different losses. Post-processing by filtering for small object removal is reported by *Bob der Baumeister*.

Four out of the top 10 approaches report the use of the MMSegmentation library even though multi-label segmentation is not natively supported. *Bridge Protector* circumvents this problem by learning 19 separate models, each covering one class. *Sheoran* later moved to segmentation-models-pytorch for finetuning. In total three out of the top 10 report to use the segmentation-models-pytorch library.

4.3. Top 3 Approaches

First Place. The winning approach by *Sheoran*[11] is based on an ensemble of the predictions of multiple models. After training using the MMSegmentation framework, the model was transferred to the segmentation-models-pytorch library for handling of the multi-label segmentation. Softmax was

[11] https://github.com/HarshitSheoran/dacl10k

			IoU [%]																		
			Concrete Defects									General Defects				Objects					
Rank	Team	mIoU [%]	Crack	ACrack	Efflorescence	Rockpocket	WConccor	Hollowareas	Cavity	Spalling	Restformwork	Wetspot	Rust	Graffiti	Weathering	ExposedR.	Bearing	EJoint	Drainage	PEquipment	JTape
									Testfinal												
1	Sheoran	**51.0**	34.1	58.2	49.9	35.3	13.6	**64.7**	22.2	**53.4**	39.9	**33.6**	52.1	**73.4**	43.8	44.6	74.4	66.3	**76.7**	82.5	49.4
2	Bridge Protector	50.5	**41.1**	**58.8**	**50.0**	30.7	10.4	62.0	22.9	51.4	**45.9**	26.7	**52.3**	73.0	42.2	46.0	76.6	66.9	71.6	**83.1**	48.4
3	Winning Wieners	50.0	35.9	50.7	48.9	28.0	16.1	63.8	19.3	50.7	42.2	31.4	**52.3**	73.0	**45.7**	**47.3**	**77.2**	63.8	71.8	82.8	**50.0**
4	Bob der Baumeister	49.7	36.6	50.4	48.3	**37.3**	**20.3**	61.7	19.4	50.2	38.9	29.8	**52.3**	71.9	43.7	42.3	73.7	64.1	75.2	79.3	48.8
5	mp269546	47.6	29.5	50.1	46.6	27.6	15.0	59.2	16.5	50.0	37.1	29.7	51.0	72.5	42.7	42.1	69.2	**68.9**	69.1	82.6	44.3
6	Shivesh Khaitan	45.7	30.4	48.9	43.8	22.6	14.2	57.4	23.2	48.3	33.4	25.9	49.0	70.1	42.0	37.4	68.7	65.3	64.7	78.0	44.4
7	Lars Nieradzik	45.1	31.2	55.5	42.9	29.5	12.5	58.4	**26.1**	46.2	31.4	27.1	47.7	66.3	40.5	34.2	67.7	58.5	62.3	74.6	44.5
8	SoloLearn	37.3	20.3	33.4	34.1	18.9	12.5	49.3	10.2	38.0	28.0	20.4	35.6	69.6	34.7	10.2	69.0	55.1	56.5	78.7	33.3
9	*dacl-squad (30 epochs)*	*37.2*	*23.5*	*38.1*	*40.1*	*19.4*	*7.8*	*46.4*	*10.6*	*42.9*	*24.7*	*20.2*	*44.7*	*62.5*	*36.0*	*33.2*	*53.9*	*49.0*	*52.1*	*69.2*	*33.1*
10	whis	36.1	15.1	25.7	36.0	19.9	11.0	45.5	14.5	43.3	35.7	8.0	3.1	45.2	36.3	33.8	62.4	61.4	71.1	77.7	40.4
									Development												
1	Sheoran	**49.8**	35.9	49.3	**46.4**	31.7	13.6	**63.5**	29.0	**53.0**	41.6	**35.5**	50.9	**71.3**	44.9	34.9	70.0	**67.5**	**78.0**	81.1	47.9
2	Winning Wieners	48.5	37.1	**49.5**	42.8	19.4	**19.6**	62.8	19.9	52.2	**41.9**	29.7	50.6	69.8	**45.4**	**39.2**	**71.0**	65.9	69.2	**82.5**	**53.4**
3	Bob der Baumeister	48.3	**38.3**	44.2	44.5	**32.4**	17.8	58.9	24.4	49.9	38.6	29.8	**51.5**	70.1	42.6	33.9	66.0	66.1	77.8	80.5	50.0
4	Lars Nieradzik	43.3	32.8	46.2	38.1	19.7	15.3	54.3	**32.1**	45.8	30.1	26.5	47.0	63.2	40.1	25.1	65.9	61.0	55.5	78.0	46.3
5	SoloLearn	35.6	18.4	25.4	29.8	22.8	14.1	40.9	24.6	35.5	28.2	18.3	31.1	64.6	40.7	20.7	58.9	44.8	51.4	73.5	32.1
6	whis	35.3	14.2	19.4	30.1	32.0	12.4	40.4	18.9	41.1	34.6	4.6	2.6	42.2	36.6	23.2	62.4	62.3	74.3	75.8	43.7
7	*dacl-squad (10 epochs)*	*34.6*	*22.8*	*34.4*	*34.8*	*12.4*	*7.4*	*45.4*	*19.7*	*41.8*	*22.2*	*20.2*	*42.2*	*54.8*	*35.6*	*32.6*	*49.4*	*45.8*	*47.1*	*54.8*	*33.8*
8	Shivesh Khaitan	31.7	13.7	33.7	27.8	8.7	8.6	45.2	14.8	37.2	18.2	19.6	29.5	55.8	33.8	24.5	49.2	47.1	43.9	61.1	30.2
9	Untitled	31.4	22.9	23.5	25.3	16.3	8.5	44.4	20.0	38.5	22.1	18.3	39.6	50.8	34.4	31.3	47.3	39.6	20.1	60.9	32.4
10	ComputingStones	31.3	24.2	25.7	32.1	5.3	8.9	37.1	21.2	34.8	24.7	16.6	36.8	48.8	30.4	12.2	53.6	46.2	40.5	60.3	36.1

Table 4. Results on the Testfinal and the Development phase. The final ranking is based on the Testfinal set.

replaced by sigmoid and JaccardLoss and BCELoss were employed. Various augmentations based on the albumentations library were applied (randomly resized crops, flips, rotations, coarse dropout, and the ImageNet normalization) complemented by CutMix [36] augmentations. Optimization was performed with the RangersLars optimizer and the scheduler CosineAnnealingLR adjusted the learning rate. The two models ConvNeXt-Large [25] and EVA-02-Large [13] pre-trained on ADE20K were trained with different initial learning rates, augmentation schemes, and heads (UPerNet and U-Net) resulting in six different models. Scoring about 40% mIoU with MMSegmentation, the EVA-02-Large with all augmentations and a UPerNet head achieved 47.8% after transfer to the segmentation-models-pytorch library. The predictions of the six models were combined for some classes, improving the overall performance.

Second Place. *Bridge Protector* also used the MMSegmentation library to train the Mask2Former [10] model with InternImage-H [33] as backbone. The pre-trained weights from ADE20K were used. In terms of augmentation, cropping and flipping were used followed by normalization. The multi-label property of the dataset is approached by decomposing the task into separate models for each category, resultng in 19 separate models. The predictions of the 19 models are concatenated.

Third Place. The *Winning Wieners*[12] combine the feature pyramid network (FPN) [21, 24] with a multi-axis vision transformer (MaxViT) [32]. FPN joins a bottom-up with a top-down pathway to handle objects across scales more effectively. MaxViT serves as a backbone, which integrates convolutional blocks based on MobileNets [17] with the attention logic from vision transformers. The xlarge MaxViT pre-trained on ImageNet was used. Standard augmentations were applied based on the RandAugment suite. The model is trained using five-fold cross-validation yielding five different models. Threshold tuning is performed for generating an ensemble on the level of predictions.

4.4. Results and Findings

The result of the dacl-challenge indicates that transfer learning on established models is a powerful tool for the domain of visual bridge inspections. Available libraries such as segmentation-models-pytorch and MMSegmentation provide good starting points for learning effective

[12] https://github.com/MIC-DKFZ/semantic_segmentation/tree/challenge/DACL

models. The drawback of MMSegmentation not providing multi-label functionality is resolved by training separate models for each class or moving to the segmentation-models-pytorch library, respectively. Training separate, class-wise models, however, induces significant computational overhead. Furthermore, ensembles appear crucial to achieve top performance by exploiting the strengths of different models. Even though achieving a good performance, the customized method from *Winning Wieners* did not exceed that of the transfer learning approaches. Apparently, the specific method configuration during training including augmentation schemes, optimizer, and other hyperparameters, are the major determinants of performance, rather than architectural considerations.

5. Conclusions

In conclusion, the dacl-challenge hosted at the “1st Workshop on Vision-Based Structural Inspections in Civil Engineering”, held during WACV 2024, marks a pivotal moment in advancing the field of computer vision for structural assessments. The participation of 23 teams underscores the interest and significance of leveraging computer vision in ensuring the safety and integrity of civil engineering structures.

The competition witnessed remarkable achievements, yet only eight teams surpassed the established baseline. Notably, the team *Sheoran* utilizing a prediction ensemble from multiple models emerged as the frontrunner, showcasing an mIoU of 51%. Their methodology sets a new benchmark in pixel-level damage recognition, laying the groundwork for future advancements in this domain.

The outcomes of this challenge highlight the potential of computer vision models in revolutionizing the inspection process for civil engineering structures. While celebrating the progress made, it also underscores the need for continued efforts in refining algorithms, enhancing dataset diversity, and exploring new avenues for more accurate, efficient, and reliable structural assessments.

As we move forward, the insights gleaned from this challenge serve as a springboard for further exploration and collaboration within the realm of vision-based structural inspections. We think that the successes and lessons learned from this will catalyze continued innovation, contributing to safer, more resilient infrastructure.

Acknowledgments. We thank Sebastian Mayr, Alexander Kern and Jürgen Hamm for their support. We are most grateful to the enterprises NetApp Deutschland GmbH and especially Ultra Consulting Network GmbH which provided the prize money. Moreover, we gratefully acknowledge the compute time granted by the Institute for Distributed Intelligent Systems and provided on the GPU cluster Monacum One at University of the Bundeswehr Munich. The challenge was supported by the MoBaP research project (IUK 639/003) financed by the Bavarian Ministry of Economic Affairs and the ev.AI.luate project as part of the FFG funding program Mobility of the Future (FO999893305).

References

[1] American Road & Transportation Builders Association (ARTBA). ARTBA bridge report, 2023. 2

[2] Haoping Bai, Gokberk Cinbis, Meng Cao, Tatiana Likhomanenko, and Shancong Mou Oncel Tuzel. 1st workshop on Vision-based InduStrial InspectiON (VISION@CVPR2023). https://vision-based-industrial-inspection.github.io/cvpr-2023/. Accessed: 2023-11-15. 3

[3] European Investment Bank. and Ipsos Public Affairs. *EIB investment survey 2023: European Union overview.* Publications Office, 2023. 2

[4] Bruno M. Belin. Rapport d’information. *Session Ordinaire*, 669, 2022. 2

[5] Christian Benz, Paul Debus, Huy Khanh Ha, and Volker Rodehorst. Crack segmentation on uas-based imagery using transfer learning. In *2019 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6, 2019. 3

[6] Christian Benz and Volker Rodehorst. Image-based detection of structural defects using hierarchical multi-scale attention. In *DAGM German Conference on Pattern Recognition (GCPR)*, pages 337–353. Springer, 2022. 3

[7] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009. Video-based Object and Event Analysis. 2

[8] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation, 2017. 3

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation, 2018. 3

[10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 7

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[12] Aaron Defazio and Samy Jelassi. Adaptivity without compromise: a momentumized, adaptive, dual averaged gradient method for stochastic optimization. *The Journal of Machine Learning Research*, 23(1):6429–6462, 2022. 6

[13] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. 7

[14] Johannes Flotzinger, Philipp J. Rösch, and Thomas Braml. dacl10k: Benchmark for semantic bridge damage segmentation, 2023. 2, 4, 5, 6

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[16] Joakim Bruslund Haurum, Mingzhu Wang, Ajmal Mian, and Thomas B. Moeslund. Computer Vision for Civil and Infrastructure Engineering Workshop (CVCIE @ ECCV2022). https://vap.aau.dk/cvcie/. Accessed: 2023-11-15. 3

[17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 7

[18] Philipp Hüthwohl, Ruodan Lu, and Ioannis Brilakis. Multi-classifier for reinforced concrete bridge defects. *Automation in Construction*, 105, Sep 2019. 3

[19] Philipp Hüthwohl and Ioannis Brilakis. Detecting healthy concrete surfaces. *Adv. Eng. Inf.*, 37:150–162, Aug. 2018. 3

[20] A. Kirillov, R. Girshick, K. He, and P. Dollar. Panoptic feature pyramid networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6392–6401, Los Alamitos, CA, USA, jun 2019. IEEE Computer Society. 3

[21] Alexander Kirillov, Kaiming He, Ross Girshick, and Piotr Dollár. A unified architecture for instance and semantic segmentation. In *CVPR*, 2017. 7

[22] Shreyas Kulkarni, Shreyas Singh, Dhananjay Balakrishnan, Siddharth Sharma, Saipraneeth Devunuri, and Sai Chowdeswara Rao Korlapati. Crackseg9k: A collection and benchmark for crack segmentation datasets and frameworks. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 179–195, Cham, 2023. Springer Nature Switzerland. 3

[23] Shengyuan Li and Xuefeng Zhao. Image-Based Concrete Crack Detection Using Convolutional Neural Network and Exhaustive Search Technique. *Advances in Civil Engineering*, 2019(MI), 2019. 3

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 7

[25] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 7

[26] Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andac Hamamci, Khan M. Iftekharuddin, Raj Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José António Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen J. Price, Tammy Riklin Raviv, Syed M. S. Reza, Michael Ryan, Duygu Sarikaya, Lawrence Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos A. Silva, Nuno Sousa, Nagesh K. Subbanna, Gabor Szekely, Thomas J. Taylor, Owen M. Thomas, Nicholas J. Tustison, Gozde Unal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koen Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. 2

[27] Guido Morgenthal, Norman Hallermann, Jens Kersten, Jakob Taraben, Paul Debus, Marcel Helmrich, and Volker Rodehorst. Framework for automated uas-based structural condition assessment of bridges. *Automation in Construction*, 97:77–95, 2019. 2

[28] Martin Mundt, Sagnik Majumder, Sreenivas Murali, Panagiotis Panetsos, and Visvanathan Ramesh. Meta-learning convolutional neural architectures for multi-target concrete defect classification with the concrete defect bridge image dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[29] Patrik F. Raudaschl, Paolo Zaffino, Gregory C. Sharp, Maria Francesca Spadea, Antong Chen, Benoit M. Dawant, Thomas Albrecht, Tobias Gass, Christoph Langguth, Marcel Lüthi, Florian Jung, Oliver Knapp, Stefan Wesarg, Richard Mannion-Haworth, Mike Bowes, Annaliese Ashman, Gwenael Guillard, Alan Brett, Graham Vincent, Mauricio Orbes-Arteaga, David Cárdenas-Peña, German Castellanos-Dominguez, Nava Aghdasi, Yangming Li, Angelique Berens, Kris Moe, Blake Hannaford, Rainer Schubert, and Karl D. Fritscher. Evaluation of segmentation methods on head and neck ct: Auto-segmentation challenge 2015. *Medical Physics*, 44(5):2020–2036, 2017. 2

[30] Timo Scharwächter, Markus Enzweiler, Uwe Franke, and Stefan Roth. Efficient multi-cue scene segmentation. In Joachim Weickert, Matthias Hein, and Bernt Schiele, editors, *Pattern Recognition*, pages 435–445, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. 2

[31] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan H. Heckers, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso.

A large annotated medical image dataset for the development and evaluation of segmentation algorithms, 2019. 2

[32] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022. 7

[33] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023. 7

[34] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers, 2021. 3, 6

[35] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642, 2020. 2

[36] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 7