Data Paper

MDMCS: A Benchmark Data Set for Multidamage Monitoring of Concrete Structures

Pengwei Guo[1]; Zhan Jiang[2]; Weina Meng[3]; and Yi Bao, Aff.M.ASCE[4]

Abstract: Concrete structures deteriorate over time due to environmental exposure and mechanical stress, leading to various types of damage such as cracking, spalling, corrosion, and exposed rebar. Automated detection using deep learning-based computer vision techniques is limited by the lack of high-quality, annotated data sets. To address this challenge, this paper presents multidamage monitoring of concrete structures (MDMCS), a data set of 1,200 images with precise pixelwise annotations involving four types of damage (cracking, spalling, corrosion, and exposed rebar) and diverse lighting conditions and material textures. The data set was evaluated using six state-of-the-art segmentation models, validating the efficacy of the data set and providing benchmarks for damage detection models. MDMCS will facilitate advances in artificial intelligence-powered structural monitoring and robot-assisted automatic inspection for improving the operation and maintenance of concrete structures. **DOI: 10.1061/JBENF2.BEENG-7893**. *© 2025 American Society of Civil Engineers*.

Author keywords: Concrete damage detection; Deep learning; Semantic segmentation; Structural health monitoring; High-resolution data set.

Introduction

Various damages that compromise the safety and durability of concrete structures (Sabouni 2023) can be caused by environmental or mechanical effects, such as extreme temperature (Kosova et al. 2025), moisture (Wang et al. 2025), freeze–thaw cycles (Liao et al. 2024), and so on (Zhai et al. 2025). Representative damages include cracking (Xiang et al. 2022), spalling (Yasmin et al. 2024), and corrosion of steel bars (Tian et al. 2023), which weaken the structural integrity and safety, potentially leading to costly repairs or catastrophic failures (Guo et al. 2022). Early detection of damage is essential for extending the lifespan of infrastructure and safeguarding public safety while minimizing operation and maintenance costs (Kong et al. 2024).

Recent advances in computer vision and deep learning have enhanced the ability to detect damage (Arafin et al. 2024; Hao and Lu 2023; Kulkarni and Sabato 2024; Luo et al. 2023). Traditional methods, such as manual inspections, are labor-intensive, prone to human errors, and limited in scalability (Tan et al. 2024). The use of in situ sensors for real-time monitoring is limited by the costs of sensor deployment and maintenance (Tan et al. 2022) and the analysis of sensor data (Liu et al. 2024). Compared with traditional methods, deep learning–based computer vision methods offer low-cost solutions applicable to structures without sensors deployed (Arafin et al. 2024).

Recent research has shown that semantic segmentation techniques can effectively detect and locate concrete damage (Arafin et al. 2022; Deng et al. 2020; Sabato et al. 2023). Various deep learning models have been developed based on advanced algorithms such as U-Net (Guo et al. 2021), MaskRCNN (Chen et al. 2022), DeepLabV3+ (Fu et al. 2021), SCHNet (Pan et al. 2020), Swin-Unet (He and Lau 2024), and SegFormer (Xie et al. 2021), which enhanced analysis efficiency. The performance of deep learning–based methods depends on the availability of large-scale, high-quality data sets with annotations. Existing data sets often suffer from low-resolution images, limited damage types (Kheradmandi and Mehranfar 2022), and inadequate annotations. Currently, there is a lack of comprehensive data sets that encompass multiple types of concrete damage.

To address these challenges, this paper presents a data set for multidamage monitoring of concrete structures (MDMCS), which comprises 1,200 annotated images with pixelwise segmentation masks for four damage types, including cracks, surface spalling, corrosion, and exposed bars. The images involve varying lighting conditions and material textures. This data set provides a valuable resource for developing and evaluating deep learning models for segmentation tasks.

The remainder of the paper highlights the differences between MDMCS and relevant data sets: section "Comparison with Existing Data Sets" presents essential information about MDMCS, section "Data Set Description" presents data collection and annotation methods and statistics and evaluates MDMCS with six deep learning algorithms, namely, FPN (Lin et al. 2017), U-Net (Guo et al. 2021), LinkNet (Chaurasia and Culurciello 2017), DeepLabV3+ (Fu et al. 2021), MaNet (Fan et al. 2020), and SegFormer (Xie et al. 2021), which are described in section "Validation." The conclusions are summarized in section "Conclusions."

Comparison with Existing Data Sets

Six publicly available data sets were considered for comparison, as presented in Table 1. CFD includes 118 images and provides pixel-level annotations for concrete crack segmentation (Shi et al. 2016).

[1]Dept. of Civil Engineering and Geosciences, Delft Univ. of Technology, Stevinweg 1, Delft 2628 CN, Netherlands; Dept. of Civil, Environmental and Ocean Engineering, Stevens Institute of Technology, Hoboken, NJ 07030. Email: pguo@stevens.edu

[2]Dept. of Civil, Environmental and Ocean Engineering, Stevens Institute of Technology, Hoboken, NJ 07030. ORCID: https://orcid.org/0009-0007-4028-0624. Email: zjiang43@stevens.edu

[3]Dept. of Civil, Environmental and Ocean Engineering, Stevens Institute of Technology, Hoboken, NJ 07030. Email: wmeng3@stevens.edu

[4]Dept. of Civil, Environmental and Ocean Engineering, Stevens Institute of Technology, Hoboken, NJ 07030 (corresponding author). ORCID: https://orcid.org/0000-0002-2766-2077. Email: yi.bao@stevens.edu

Note. This manuscript was submitted on June 17, 2025; approved on October 9, 2025; published online on December 11, 2025. Discussion period open until May 11, 2026; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Bridge Engineering***, © ASCE, ISSN 1084-0702.

Table 1. Statistical metadata

Annotation type	Data set type	Label count	Image count	Pixel count (×1,000)	Pixel per image (%)
Cracking	Training	600	175	2,216	0.29
	Validation	31	15	202	0.26
	Testing	28	12	261	0.34
Spalling	Training	1,371	864	141,255	18.16
	Validation	109	82	10,985	15.13
	Testing	82	74	13,093	16.83
Corrosion	Training	703	286	11,382	1.46
	Validation	40	25	1,216	1.57
	Testing	18	15	1,311	1.69
Exposed rebar	Training	1,404	464	12,269	1.58
	Validation	100	41	734	1.25
	Testing	84	48	1,290	1.66
Background	Training	N.A.	N.A.	610,480	78.51
	Validation	N.A.	N.A.	64,622	81.79
	Testing	N.A.	N.A.	61,805	79.48
Total	Training	4,708	1,789	777,600	100
	Validation	280	163	77,760	100
	Testing	212	149	77,760	100

Note: The total image count exceeds 1,200 since some images involve multiple damage types and are therefore double-counted.

CrackTree200 includes 200 images with annotated concrete crack masks, focusing on treelike crack patterns (Zou et al. 2012). Crack500 includes 500 images of varying sizes and provides pixel-level annotations for concrete crack segmentation (Yang et al. 2020). GAPs384 includes 1,969 images for pavement distress detection (Eisenbach et al. 2017). These four data sets contain only crack images, rather than multiple types of damage relevant for bridge inspection. COCO-BRIDGE includes 1,470 images for crack and corrosion inspection, each with bounding box annotations (Bianchi et al. 2021). CODEBRIM includes 1,590 images with 5,354 annotated defect bounding boxes for five types of concrete damage: crack, spalling, exposed rebar, corrosion, and efflorescence. The COCO-BRIDGE and CODEBRIM data sets only support image-level annotation (bounding box) intended for damage detection tasks, rather than instance segmentation, as they do not provide fine-grained pixel-level masks (Mundt et al. 2019). Dacl10K comprises 10,000 images featuring 12 types of concrete damage, such as cracking, spalling, corrosion, and weathering (Flotzinger et al. 2024). Although Dacl10K focuses on damage segmentation, the labels are coarse (pixel level) and lack sufficient precision for damage segmentation, leading to limited accuracy of machine learning models. For example, the mean intersection over union (IoU) was 0.424, and the specific IoU values were 0.286 for cracking and 0.406 for spalling. In short, existing data sets are limited in their ability to support multidamage segmentation tasks with fine-grained pixel-level masks.

Data Set Description

The presented data set comprises a total of 1,200 images, with 1,000 allocated for training, 100 for validation, and 100 for testing. The training set is used to optimize the model parameters during the learning process. The validation set is employed to monitor the performance of models after each training epoch, helping to tune hyperparameters and prevent overfitting. The model that achieves the highest accuracy on the validation set is selected as the best-performing model. Finally, the testing set, which remains separate from the training and validation phases, is used to assess the generalization ability of models on unseen data. The resolution of the images is 1,080 × 720 pixels, which were collected from various types of bridges, including highway, railway, and pedestrian bridges, sourced from both field collections and image repositories. The annotation format is pixelwise segmentation mask. The data set was constructed in four steps, as summarized in Fig. 1:

1. *Image collection*: High-resolution images were collected from concrete structures via multiple inspection projects involving various material textures, supplemented by online images. The images were captured in 2019 using the front camera of an iPhone X, which has a 7-megapixel sensor and a maximum resolution of 3,264 × 2,448 pixels. The images were manually cropped to highlight the regions of interest and achieved a consistent size.
2. *Image annotation*: The images were annotated using the LabelMe toolbox for pixel-level segmentation of various damage types with fine-grained annotation. Each annotation corresponds to clearly defined damage regions such as cracks, spalling, and rebar exposure, ensuring high-quality masks suitable for segmentation tasks. The labeling process was carried out manually by tracing the boundaries of each damaged area through point-by-point clicking. The coordinates of the clicked points were saved in JSON files, which were then converted into corresponding mask images. Representative examples of images and their annotations are available in section "Image Annotation."
3. *Image augmentation*: The original data set involved an imbalanced data distribution among the different damage types. The data set includes more images of spalling and cracks, with fewer of corrosion and rebar exposure. To mitigate data imbalance, geometric transformations, such as rotation, flipping,

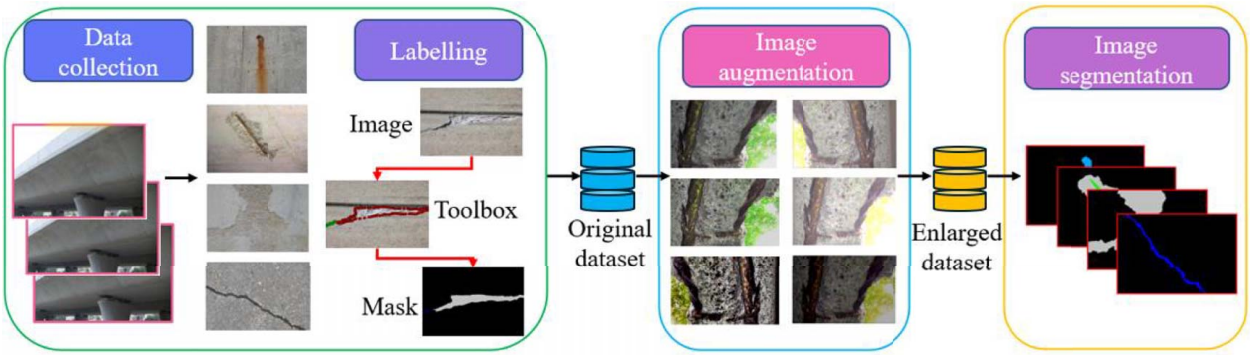

Fig. 1. General workflow of this research.

scaling, and cropping, were applied, along with Gaussian noise addition and brightness adjustments to create variations in underrepresented damage. The augmentation was performed using the Albumentations library in Python (Buslaev et al. 2020). More details are available in section "Data Augmentation."

4. *Image segmentation validation*: The performance of the data set was evaluated using six deep learning models for multidamage semantic segmentation tasks. This information is elaborated in section "Results."

The statistical metadata is summarized in Table 1, including the label counts, image count, pixel count, and pixel percentage for the training, validation, and testing sets. The data set includes various types of concrete damage, such as cracking, spalling, corrosion, and rebar exposure. Cracking occurs more frequently but covers a much smaller pixel area, highlighting its fine-grained nature. Corrosion and exposed rebar are less represented. Background pixels dominate (approximately 78%–82%), reflecting the typical imbalance in bridge inspection images.

The variation of lighting conditions of images is shown in Fig. 2. The average pixel intensity of each image exhibits probability distributions reflecting the existence of images captured from low-light or overexposure conditions. According to Guo et al. (2024b), an average intensity lower than 30 suggests low light, and an average intensity higher than 225 indicates overexposure. An average intensity between 30 and 225 means normal lighting conditions.

Validation

Evaluation Method

Six semantic segmentation architectures were selected based on the architectural diversity: (1) U-Net, a popular encoder–decoder model for structural segmentation tasks (Ronneberger et al. 2015); (2) LinkNet, a popular model utilizing an encoder–decoder architecture with additive skip connections to preserve spatial details while enhancing computational efficiency (Chaurasia and Culurciello 2017); (3) feature pyramid network (FPN), which enhances multiscale feature learning (Lin et al. 2017); (4) DeepLabV3+, a segmentation model utilizing atrous spatial pyramid pooling to capture multiscale contextual information and employs a decoder module to refine segmentation results (Chen et al. 2018); (5) MaNet is a segmentation model that enhances U-Net by incorporating multiscale attention mechanisms to improve feature extraction and segmentation accuracy (Fan et al. 2020); and (6) SegFormer, a transformer-based segmentation model that effectively balances efficiency and accuracy (Xie et al. 2021). The selected models vary significantly in complexity, ranging from lightweight architectures like FPN and U-Net to more advanced and computationally intensive models such as MaNet and Segformer. This range enables a balanced evaluation of tradeoffs between model performance and computational efficiency. Table 2 summarizes the model size and inference speed based on 256 × 256 input images for a comparison of model complexity.

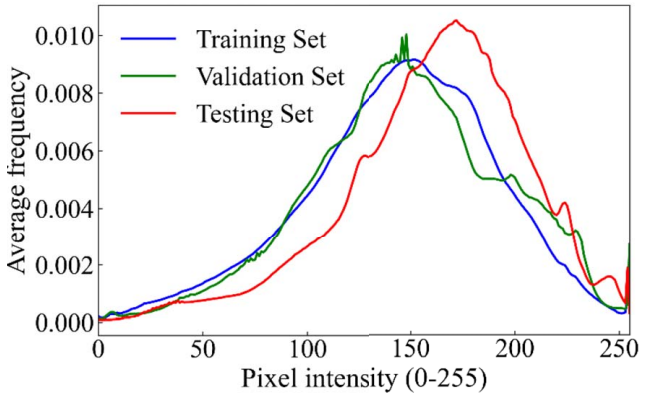

Fig. 2. Distribution of average pixel intensity across training, validation, and testing data sets.

The accuracy of the models was evaluated using performance metrics elaborated in section "Performance Metrics." After the most accurate model was selected, further evaluation was carried out to enhance its performance via leveraging data augmentation techniques. All experiments were conducted on an RTX 4090 GPU using PyTorch Lightning 2.5.0 and CUDA 11.8, leveraging high computational power for efficient training and evaluation.

Performance Metrics

The validity of the data set was evaluated by assessing the accuracy of segmentation models trained using the data set. These metrics include mean pixel accuracy (mPA), mean IoU (mIoU), and dice score (F1 score), providing a comprehensive analysis of model performance. mPA, mIoU, and mDice range from 0 to 1.

mPA is a performance metric used in semantic segmentation to evaluate the proportion of correctly classified pixels across all classes. Unlike standard pixel accuracy, mPA calculates the accuracy for each class separately and then averages them, ensuring that smaller or less frequent classes are not overshadowed by dominant ones. The mPA is defined in the following equation (Yu et al. 2023):

$$\text{mPA} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i + \text{TN}_i}{\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i} \tag{1}$$

For multiclass damage segmentation, IoU evaluates the overlap between predicted and ground truth segmentation masks. To account for multiple damage classes, mIoU is calculated by averaging the IoU scores across all classes, ensuring a balanced evaluation of segmentation performance across different types of damage. The mIoU is defined in the following equation (Dais et al. 2021):

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i} \tag{2}$$

For multiclass damage segmentation, the mean dice coefficient (mDice), also known as the F1 score for segmentation, measures the similarity between predicted and ground truth masks. It is calculated as the average dice score across all classes. mDice is

Table 2. Comparison between various segmentation models

Model	Model parameters (M)	Inference speed (ms)
FPN	23.6	41.8
U-Net	29.0	49.5
LinkNet	32.7	59.1
DeepLabV3+	45.1	64.0
SegFormer	64.1	69.7
MaNet	146.9	97.2

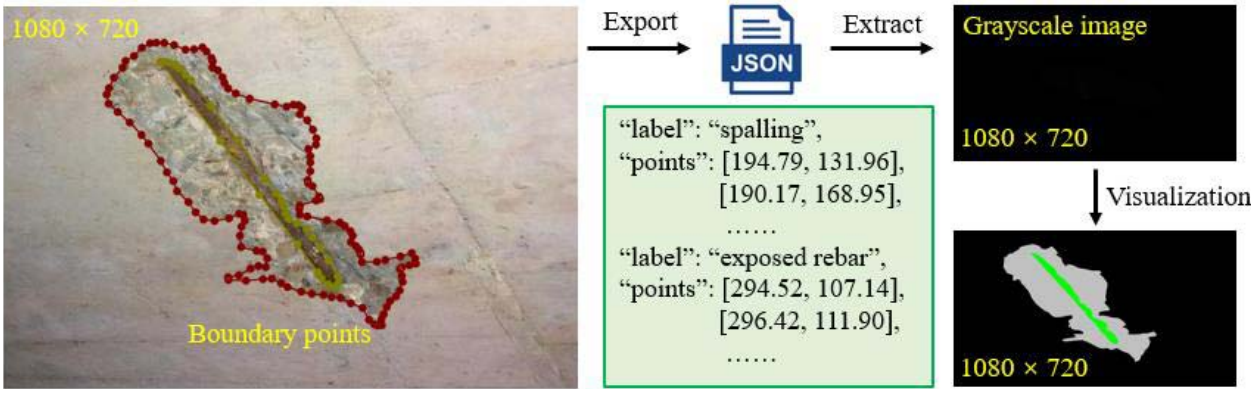

Fig. 3. Illustration of mask generation using the LabelMe toolbox.

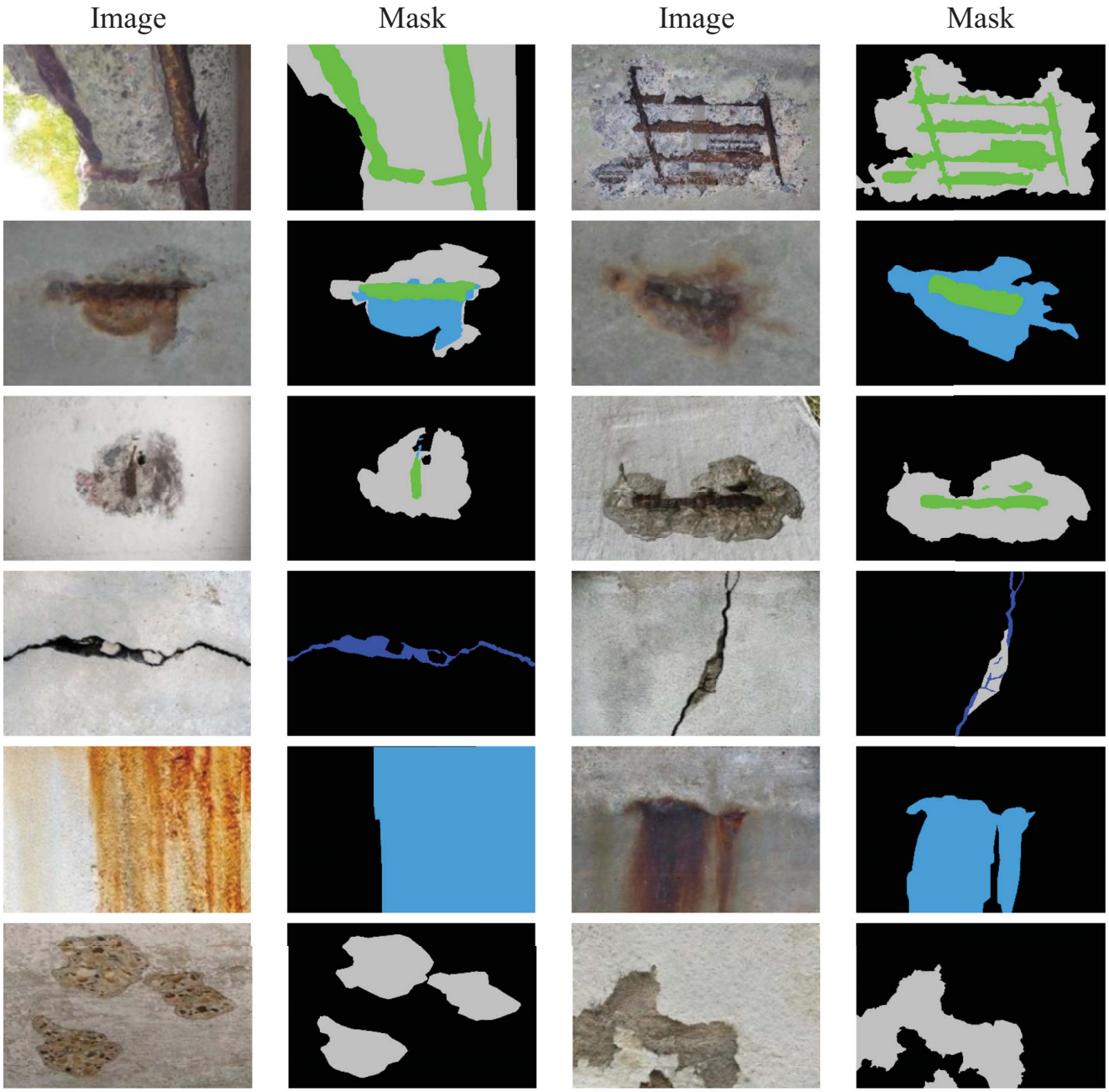

Fig. 4. Representative examples of established data sets.

defined in the following equation (Guo et al. 2024b):

$$\text{mDice} = \frac{1}{N}\sum_{i=1}^{N}\frac{2\times \text{TP}_i}{2\times \text{TP}_i + \text{FP}_i + \text{FN}_i} \tag{3}$$

where N ($N = 4$) = total number of classes; $i = i$th class; and TP, FP, and FN = true positive, false positive, and false negative, respectively. TP occurs when positive instances (i.e., cracks) are correctly identified, FP occurs when uncracked concrete is incorrectly identified as cracked, and FN occurs when cracks are incorrectly identified as uncracked concrete.

Image Annotation

The image labels were initially converted into JSON format using the LabelMe toolbox, which contains details such as the boundary point coordinates and damage class. As shown in Fig. 3, the JSON file stores the boundary point coordinates of spalling and exposed rebar in a structured format. The JSON files were then processed and transformed into grayscale segmentation masks, where each pixel value corresponds to a specific class. In the grayscale image, pixel values were assigned as follows: 0 for background, 1 for crack, 2 for spalling, 3 for corrosion, and 4 for exposed rebar. This transformation enables the data set to be efficiently used for multiclass damage segmentation, ensuring that each damage type is clearly distinguished for model training and evaluation. For visualization purposes, the 0–4 values were mapped to the following colors: (0, 0, 0), (0, 0, 255), (192, 192, 192), (0, 162, 255), and (0, 255, 0), respectively.

Fig. 4 provides a visualization of the data set, showcasing representative image–mask pairs to illustrate the variety and annotation quality of the samples. These examples indicate that fine-grained pixel-level masks have been generated for various damage types.

Data Augmentation

The data augmentation pipeline randomly applies a combination of operations to each training image, including horizontal flipping, padding, cropping, Gaussian noise, and perspective transformation,

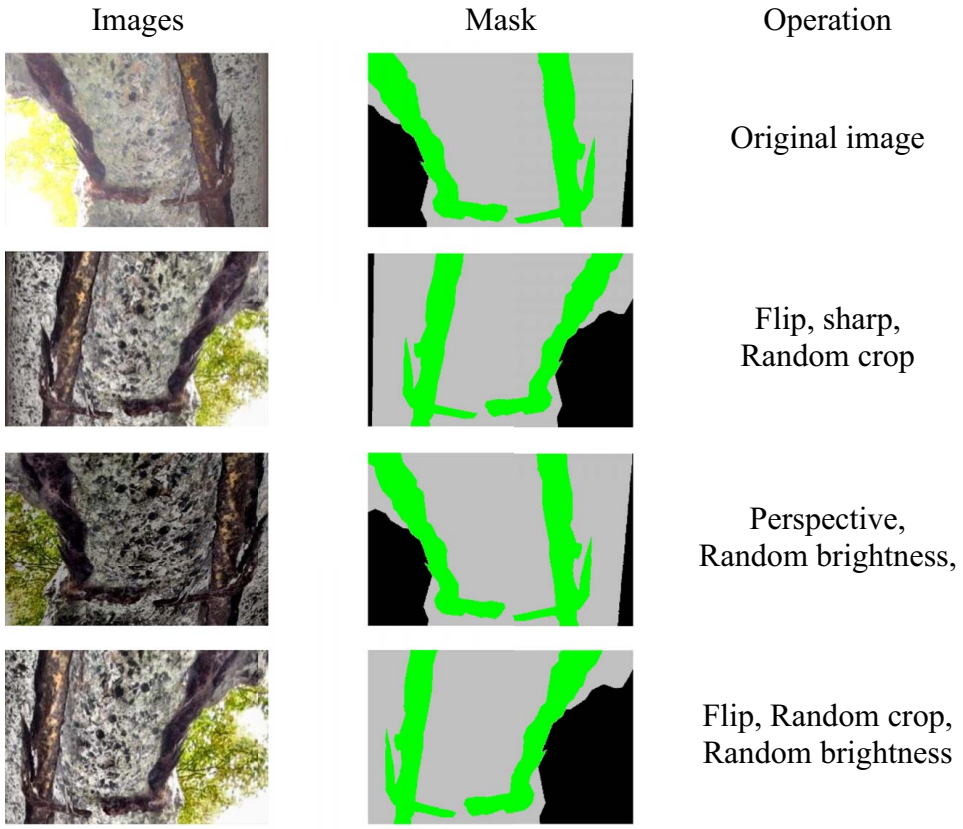

Fig. 5. Representative examples of data augmentation operations.

Table 3. Segmentation accuracy of trained models on different data sets

Data set	Training			Validation			Testing		
Model	mIoU	mDice	mPA	mIoU	mDice	mPA	mIoU	mDice	mPA
FPN	0.808	0.893	0.988	0.763	0.859	0.992	0.746	0.847	0.990
U-Net	0.821	0.897	0.985	0.757	0.855	0.991	0.808	0.891	0.990
LinkNet	0.813	0.877	0.984	0.760	0.857	0.991	0.813	0.895	0.991
DeepLabV3+	0.825	0.867	0.989	0.766	0.861	0.992	0.818	0.899	0.993
MaNet	0.720	0.837	0.977	0.745	0.845	0.990	0.803	0.890	0.990
SegFormer	0.766	0.867	0.985	0.747	0.849	0.990	0.746	0.848	0.990

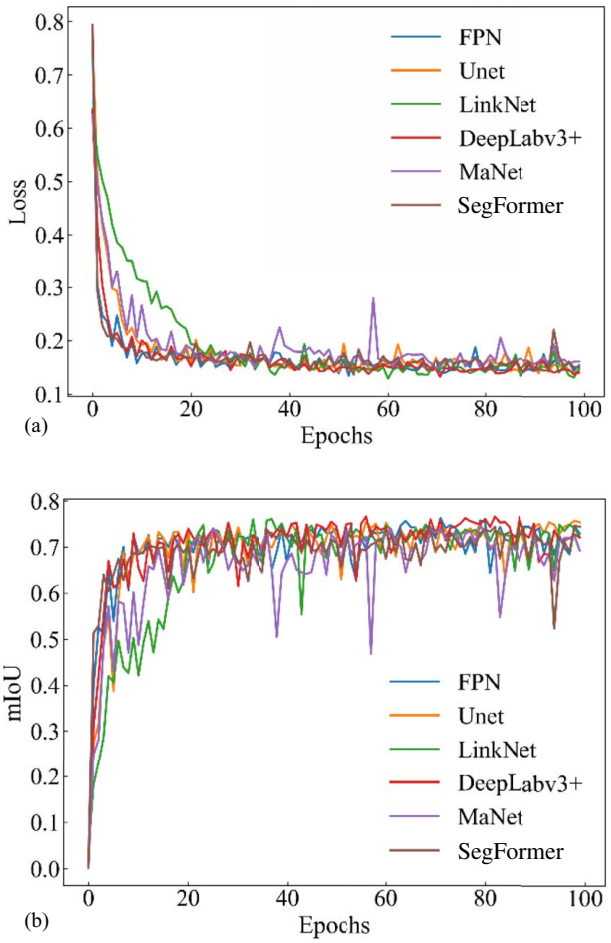

Fig. 6. Convergence analysis results: (a) loss; and (b) mIoU.

brightness/contrast adjustment, sharpening, and blurring, ensuring diverse and realistic variations for robust model training. These transformations help simulate various real-world conditions such as lighting changes, motion artifacts, and sensor noise, which enhances the generalization capability of models. By introducing such randomness and variability, the augmented data set effectively mitigates overfitting and improves the performance of the segmentation model on unseen data. Data augmentation was not applied to the validation and test sets to maintain an unbiased evaluation. Representative examples of the applied augmentation operations are presented in Fig. 5. The images were modified through data augmentation, and the corresponding masks underwent the same geometric transformations to maintain pixelwise alignment. The augmentation was specifically applied to underrepresented classes, such as exposed rebar and corrosion, to address class imbalance issues. The detailed discussion of the effects of data augmentation can be found in section "Segmentation Performance."

Results

Segmentation Performance

For the baseline model evaluation, a learning rate of 2×10^{-4}, a batch size of 16, and an input resolution of 640×480 were used with the Adam optimizer. To maintain pixel integrity, the image was not resized, and random cropping was applied to achieve the target size. Each model was trained for 100 epochs to identify the best one that achieves the highest mIoU for evaluation on the test data set. Table 3 presents the segmentation accuracy of different models across the training, validation, and testing data sets, evaluated using mIoU, mDice, and mPA.

Among all models, DeepLabV3+ achieved the highest testing performance, with mIoU = 0.818, mDice = 0.899, and mPA = 0.993, demonstrating its superior segmentation capability. LinkNet followed closely, attaining an mIoU of 0.813, while U-Net and

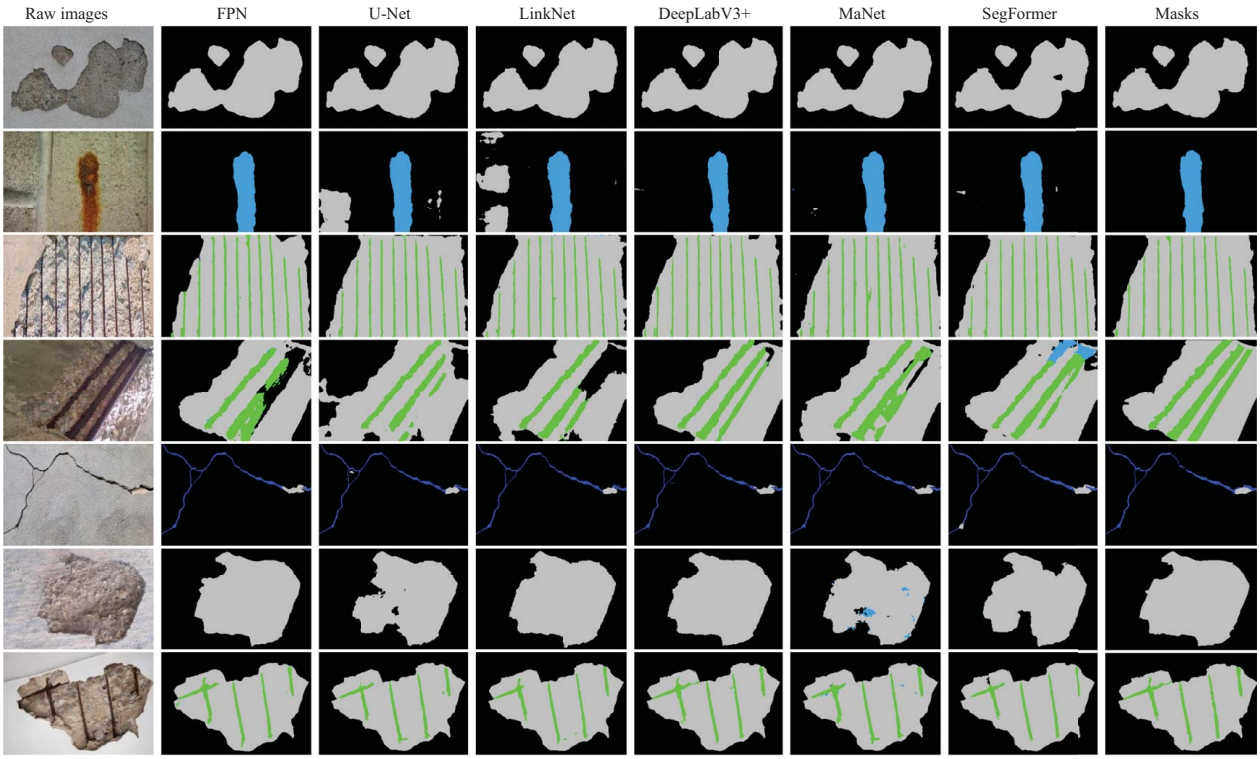

Fig. 7. Comparison of segmentation performance across different semantic segmentation models.

Table 4. Impact of data augmentation strategies on segmentation performance

	Without AG		AG		Partial AG	
Damage type	Training	Testing	Training	Testing	Training	Testing
Crack	—	0.846	—	0.850	—	0.858
Spalling	—	0.868	—	0.908	—	0.904
Corrosion	—	0.817	—	0.808	—	0.826
Exposed rebar	—	0.742	—	0.753	—	0.751
Mean	0.825	0.818	0.841	0.830	0.843	0.835

MaNet achieved mIoU scores of 0.808 and 0.803, respectively. In contrast, FPN and SegFormer recorded the lowest mIoU of 0.746. The overall results indicate that DeepLabV3+ consistently outperforms the other models across all data sets, particularly excelling in the validation and testing phases. This finding highlights its robust feature extraction and fusion capabilities. Additionally, the high segmentation accuracy across multiple models establishes it as a benchmark for damage segmentation.

The convergence analysis is shown in Fig. 6. The steady decrease in loss and the increase in accuracy over epochs indicate efficient learning. All models start with high loss values, which decrease within the first few epochs. DeepLabV3+ and SegFormer converge the fastest, as their loss values fall below 0.25 by Epoch 4, while their accuracy surpasses 0.60. Accuracy improves rapidly, with these two models exceeding 0.50 by Epoch 2; FPN, MaNet, and U-Net exhibit steadier but slower progress, while LinkNet shows the slowest convergence, suggesting weaker feature learning.

More segmentation results are presented in Fig. 7, maintaining its original resolution of 1,080 × 720 during testing. Images captured under low-light conditions show poor segmentation performance, suggesting that it is important to apply image preprocessing techniques to enhance data quality and improve the segmentation performance of deep learning models (Guo et al. 2024b).

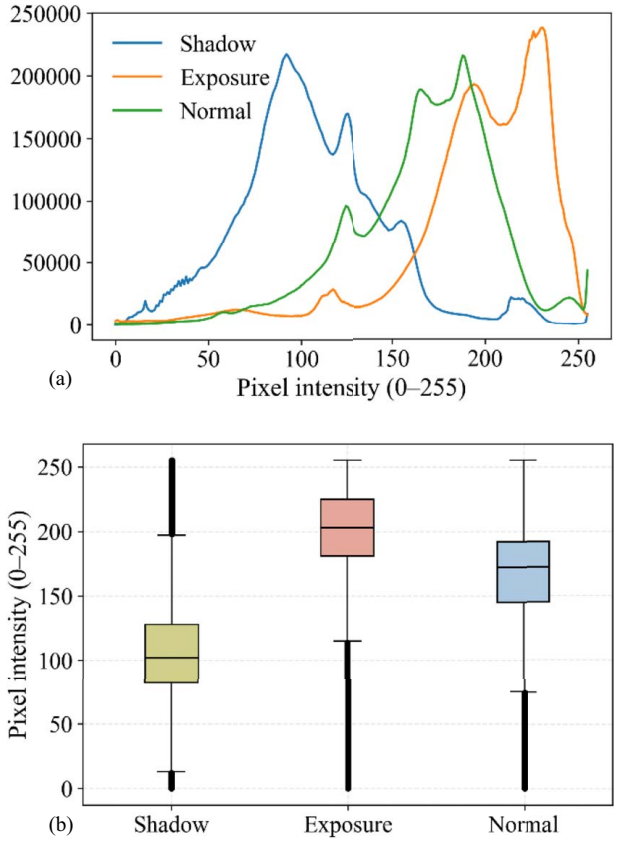

Fig. 8. Grayscale intensity analysis of images under different lighting conditions: (a) grayscale histogram distribution; and (b) boxplot comparison of pixel intensity.

Augmentation Strategies

Table 4 presents the impact of data augmentation on improving the accuracy of DeepLabV3+: (1) without AG: no augmentation. (2) AG: augmentation for all classes. The 1,000 images from the training set were augmented via random geometric transformations, brightness adjustments, and noise processing. The augmentation operation was applied to all four damage categories over 100 epochs. (3) Partial AG: augmentation for minority classes (crack and corrosion). Partial augmentation was applied to generate more images of corrosion and cracks, while the images of spalling and exposed rebar were kept. To assess generalization, the model was evaluated on a testing set composed exclusively of nonaugmented images, which reflects its ability to perform well on real-world data distributions. There is no significant difference between the training accuracy and the test accuracy, indicating that the model does not suffer from overfitting. The evaluation results show that performing partial AG achieves the highest mIoU (0.835), highlighting the effectiveness of targeted augmentation for enhancing underrepresented classes.

Generalization

This section presents a comparison of grayscale characteristics across three lighting conditions—shadow, exposure, and normal—using 20 images from the testing set for each category. Fig. 8(a) illustrates the grayscale intensity distribution across the three classes. The shadow class peaks at an intensity of 92, indicating predominantly dark regions. The normal class peaks at 188, reflecting moderate brightness. In contrast, the exposure class peaks at an intensity of 231, indicating a strong shift toward high brightness. This progressive shift from shadow to exposure highlights distinct illumination conditions. In Fig. 8(b), the grayscale intensity analysis shows that the shadow class exhibits a lower intensity range, with a median of 102, indicating predominantly darker images. The exposure class shows significantly higher intensity values, with a median of 203, reflecting brighter overall brightness. Normal class falls in between, with a median of 172. These distinct distributions highlight clear differences in brightness levels across the three categories.

The segmentation performance of the images was evaluated using the trained DeepLabV3+ model. The model achieved mIoU values of 0.855 for the exposure class, 0.847 for the normal class, and 0.775 for the shadow class. The relatively higher accuracy for exposure and normal classes can be attributed to well-lit conditions, which provide clearer feature boundaries. As long as the image is not overexposed, the quality remains unaffected. The lower performance in the shadow class may be attributed to reduced contrast and visibility under low-light conditions, which

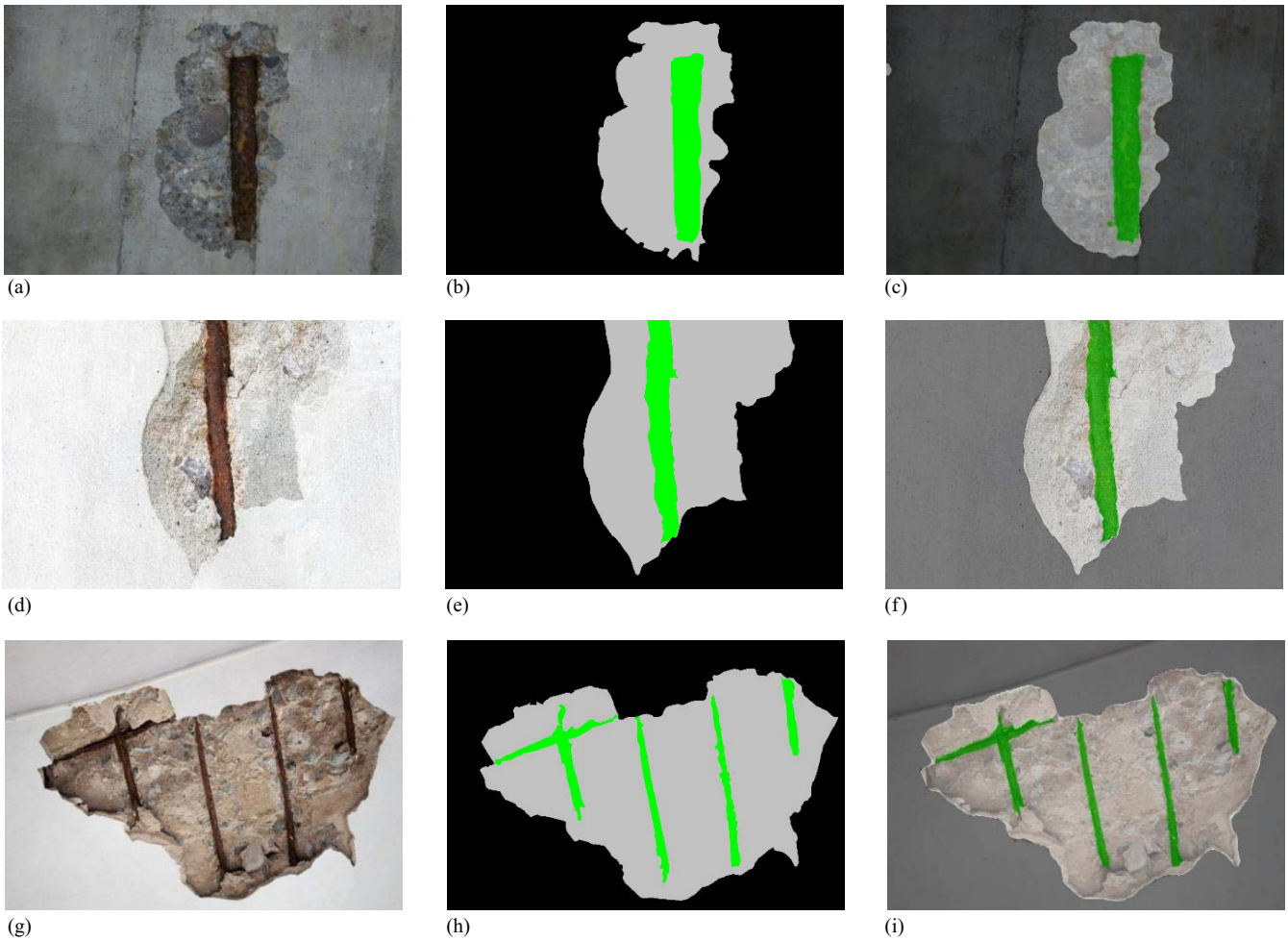

Fig. 9. Representative examples of segmentation results: (a–c) shadow condition; (d–f) exposure condition; and (g–i) normal condition. (a, d, and g) original image, (b, e, and h) ground truth mask, and (c, f, and i) model prediction.

makes feature extraction and boundary delineation more challenging. Representative examples for images under different lighting conditions are shown in Fig. 9. Boundary prediction is generally less accurate under shadow conditions.

Potential Opportunities

The MDMCS data set offers diverse opportunities for reuse in structural health monitoring and the automated inspection of civil infrastructure (Kumar et al. 2021). It is promising to reuse the data set to develop advanced segmentation or classification models that can be used in multidamage monitoring or inspection tasks for the aging infrastructure (Azimi et al. 2020). The monitoring or inspection outcomes can then be utilized to optimize asset operation and maintenance schemes (Mohamed et al. 2020). With advances in robotics, it is promising to train robots using the MDMCS data set for the automatic inspection of damage, aiming to identify and characterize infrastructure damage early or in posthazard surveys (Liu et al. 2022).

This research shows that the data set can be utilized to develop multidamage segmentation models for the automatic analysis of images, and data augmentation can enhance the accuracy of segmentation tasks. However, the performance improvement is incremental because the adopted augmentation methods inherently preserve the primary patterns and characteristics of the original images and do not produce new variations in image features. The current prediction accuracy is 0.835, indicating opportunities for further improvement. It is promising to apply or develop advanced techniques such as generative adversarial network (GAN)-based methods for data augmentation, aiming to synthesize new images with improved diversity, as elaborated in recent research on the generation of images for concrete cracks (Guo et al. 2024a; Duan et al. 2025) and restoration of low-quality images (Guo et al. 2024b). GAN-based methods have been used to augment data sets with crack images, whereas augmentation for other defect types has received limited attention.

The MDMCS data set has the potential to support cross-domain testing. Although MDMCS focuses on bridge damage, it includes a diverse range of damage types, such as cracking, spalling, corrosion, and exposed bars, across various structures and surface textures. These variations make it a suitable candidate for models intended for broader inspection tasks. For instance, an MDMCS-trained model can be evaluated on data sets involving pavements, tunnel linings, or steel surfaces to assess its generalization across different domains. To improve cross-domain performance, transfer learning can be employed by fine-tuning the MDMCS-trained model on a small set of labeled samples from the target domain. This approach leverages the learned feature representations from MDMCS and adapts them to new conditions with minimal annotation effort. Transfer learning has shown strong potential in enhancing generalization, especially when the differences between domains are small yet significant, such

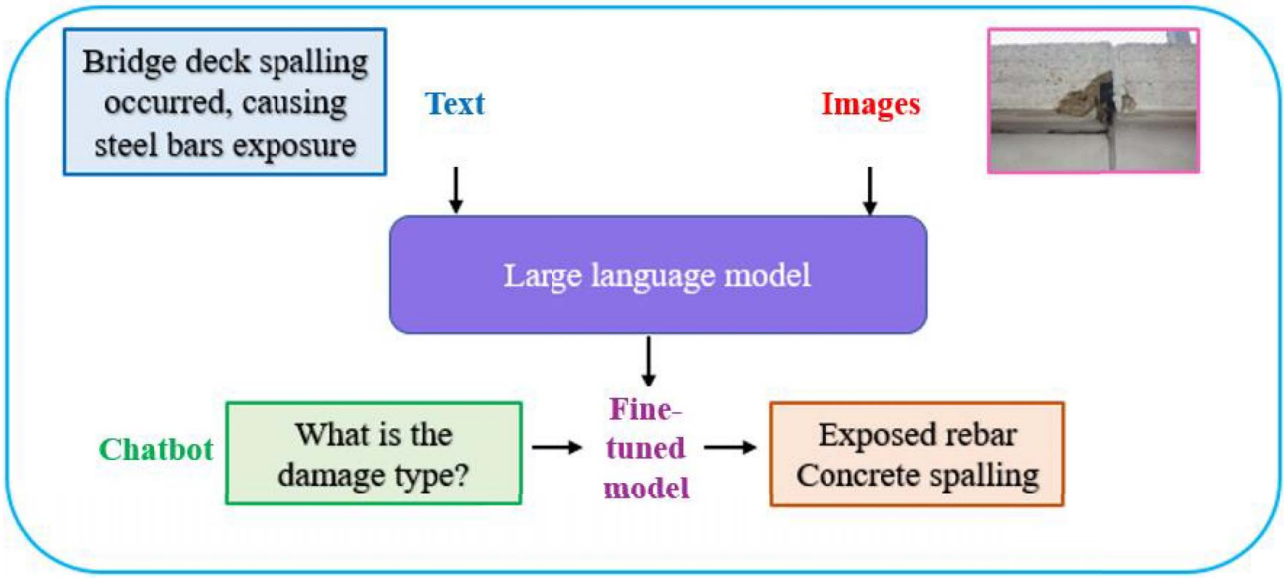

Fig. 10. Intelligent chatbot system using a fine-tuned LLM.

as changes in texture, lighting, or surface damage (Dais et al. 2021).

In future research, the data set will be further expanded by incorporating a broader range of damage scenarios, such as concrete weathering, seepage, and other relevant types, while also increasing the sample size of each damage class. In addition, the data set will incorporate high-quality text–image pairs, where each image is accompanied by detailed textual descriptions of the observed damage. This multimodal format will support the development of advanced AI applications, such as text-to-image generation using the stable diffusion model to synthesize realistic damage scenarios. It will also enable vision-language tasks with models like contrastive language-image pretraining, facilitating automated visual question answering and damage interpretation. Moreover, fine-tuning large language models with domain-specific data will help build intelligent chatbot systems capable of interaction, allowing users to query damage conditions, receive visual summaries, or obtain diagnostic guidance, as shown in Fig. 10.

Conclusions

This paper presents a data set called MDMCS, which includes 1,200 images with a resolution of 1,080 × 720 pixels. The images were labeled with precise pixelwise annotations involving four types of concrete damage, namely, cracking, spalling, corrosion, and exposed rebar, under diverse lighting and material texture conditions, aiming to mitigate the challenge of insufficient data for developing deep learning-based multidamage assessment models. The data set was evaluated using six state-of-the-art segmentation models: FPN, U-Net, LinkNet, DeepLabV3+, MaNet, and SegFormer, which validated the efficacy of the data set and provided a benchmark for damage detection models. It was found that the imbalance of different concrete damage types in the data set was alleviated by data augmentation, which improved the accuracy of the segmentation models. For future research, it is promising to utilize the MDMCS data set to develop advanced augmentation techniques and high-fidelity segmentation models that will advance artificial intelligence-powered structural monitoring and robot-assisted automatic inspection for improving the operation and maintenance of concrete structures. Future work will focus on expanding the data set to include more diverse damage types and detailed text–image pairs to support advanced AI applications, including text-to-image generation, vision-language tasks, and intelligent chatbot systems for intuitive damage analysis and interaction.

Data Availability Statement

Some or all data, models, or code generated or used during the study are available in a repository online in accordance with funder data retention policies: (Guo and Bao 2025).

Acknowledgments

This research was funded by the United States Department of Transportation through PHMSA (Grant Nos. 693JK31950008CAAP and 693JK32310008POTA) and the United States National Science Foundation (Grant No. 2305882).

Author Contributions

Pengwei Guo: Data curation, Formal analysis, Investigation, Software, Visualization, Writing—original draft; Zhan Jiang: Data curation, Investigation, Software, Validation, Writing—review and editing; Weina Meng: Conceptualization, Formal analysis, Methodology, Project administration, Resources, Validation, Writing—review and editing; Yi Bao: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Supervision, Validation, Writing—review and editing.

References

Arafin, P., A. M. Billah, and A. Issa. 2024. "Deep learning-based concrete defects classification and detection using semantic segmentation." *Struct. Health Monit.* 23 (1): 383–409. https://doi.org/10.1177/14759217231168212.

Arafin, P., A. Issa, and A. H. M. M. Billah. 2022. "Performance comparison of multiple convolutional neural networks for concrete defects classification." *Sensors* 22 (22): 8714. https://doi.org/10.3390/s22228714.

Azimi, M., A. D. Eslamlou, and G. Pekcan. 2020. "Data-driven structural health monitoring and damage detection through deep learning: State-of-the-art review." *Sensors* 20 (10): 2778. https://doi.org/10.3390/s20102778.

Bianchi, E., A. L. Abbott, P. Tokekar, and M. Hebdon. 2021. "COCO-bridge: Structural detail data set for bridge inspections." *J. Comput. Civil Eng.* 35 (3): 04021003. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000949.

Buslaev, A., V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin. 2020. "Albumentations: fast and flexible image augmentations." *Information* 11 (2): 125. https://doi.org/10.3390/info11020125.

Chaurasia, A., and E. Culurciello. 2017. "LinkNet: Exploiting encoder representations for efficient semantic segmentation." In *Proc., 2017 IEEE Visual Communications and Image Processing*, 1–4. Piscataway, NJ: IEEE.

Chen, G., S. Teng, M. Lin, X. Yang, and X. Sun. 2022. "Crack detection based on generative adversarial networks and deep learning." *KSCE J. Civ. Eng.* 26 (4): 1803–1816. https://doi.org/10.1007/s12205-022-0518-2.

Chen, L.-C., Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. 2018. "Encoder-decoder with atrous separable convolution for semantic Image segmentation." In *Proc., European Conf. on Computer Vision*, 801–818. Cham, Switzerland: Springer.

Dais, D., İ. E. Bal, E. Smyrou, and V. Sarhosis. 2021. "Automatic crack classification and segmentation on masonry surfaces using convolutional neural networks and transfer learning." *Autom. Constr.* 125: 103606. https://doi.org/10.1016/j.autcon.2021.103606.

Deng, L., H.-H. Chu, P. Shi, W. Wang, and X. Kong. 2020. "Region-based CNN method with deformable modules for visually classifying concrete cracks." *Appl. Sci.* 10 (7): 2528. https://doi.org/10.3390/app10072528.

Duan, S., X. Tan, P. Guo, Y. Guo, and Y. Bao. 2025. "The transformative roles of generative artificial intelligence in vision techniques for structural health monitoring: A state-of-the-art review." *Adv. Eng. Inf.* 68: 103719. https://doi.org/10.1016/j.aei.2025.103719.

Eisenbach, M., R. Stricker, D. Seichter, K. Amende, K. Debes, M. Sesselmann, D. Ebersbach, U. Stoeckert, and H.-M. Gross. 2017. "How to get pavement distress detection ready for deep learning? A systematic approach." In *Proc., 2017 Int. Joint Conf. on Neural Networks*, 2039–2047. Piscataway, NJ: IEEE.

Fan, T., G. Wang, Y. Li, and H. Wang. 2020. "MA-Net: A multi-scale attention network for liver and tumor segmentation." *IEEE Access* 8: 179656–179665. https://doi.org/10.1109/ACCESS.2020.3025372.

Flotzinger, J., P. J. Rösch, and T. Braml. 2024. "dacl10k: Benchmark for semantic bridge damage segmentation." In *Proc., IEEE/CVF Winter Conf. on Applications of Computer Vision*, 8626–8635. Piscataway, NJ: IEEE.

Fu, H., D. Meng, W. Li, and Y. Wang. 2021. "Bridge crack semantic segmentation based on improved Deeplabv3+." *J. Mar. Sci. Eng.* 9 (6): 671. https://doi.org/10.3390/jmse9060671.

Guo, P., W. Meng, and Y. Bao. 2021. "Automatic identification and quantification of dense microcracks in high-performance fiber-reinforced cementitious composites through deep learning-based computer vision." *Cem. Concr. Res.* 148: 106532. https://doi.org/10.1016/j.cemconres.2021.106532.

Guo, P., W. Meng, and Y. Bao. 2024a. "Intelligent characterization of complex cracks in strain-hardening cementitious composites based on generative computer vision." *Constr. Build. Mater.* 411: 134812. https://doi.org/10.1016/j.conbuildmat.2023.134812.

Guo, P., X. Meng, W. Meng, and Y. Bao. 2022. "Monitoring and automatic characterization of cracks in strain-hardening cementitious composite (SHCC) through intelligent interpretation of photos." *Composites, Part B* 242: 110096. https://doi.org/10.1016/j.compositesb.2022.110096.

Guo, P., X. Meng, W. Meng, and Y. Bao. 2024b. "Automatic assessment of concrete cracks in low-light, overexposed, and blurred images restored using a generative AI approach." *Autom. Constr.* 168: 105787. https://doi.org/10.1016/j.autcon.2024.105787.

Guo, P., and Y. Bao. 2025. "MDMCS: A benchmark dataset for multi-damage monitoring of concrete structures." *Mendeley data*. https://doi.org/10.17632/6x4dzzrs2h.2.

Hao, Z., and C. Lu. 2023. "User-friendly end-to-end fiber identification for fiber-reinforced cementitious composites (FRCC) through deep learning." *Constr. Build. Mater.* 403: 133169. https://doi.org/10.1016/j.conbuildmat.2023.133169.

He, M., and T. L. Lau. 2024. "A novel network fusing transformer and CNN for road crack segmentation." *IEEE Access* 12: 165610–165625. https://doi.org/10.1109/ACCESS.2024.3492193.

Kheradmandi, N., and V. Mehranfar. 2022. "A critical review and comparative study on image segmentation-based techniques for pavement crack detection." *Constr. Build. Mater.* 321: 126162. https://doi.org/10.1016/j.conbuildmat.2021.126162.

Kong, X., Z. Liu, H. Liu, J. Hu, and L. Deng. 2024. "Recent advances on inspection, monitoring, and assessment of bridge cables." *Autom. Constr.* 168: 105767. https://doi.org/10.1016/j.autcon.2024.105767.

Kosova, F., Ö. Altay, and H. Ö. Ünver. 2025. "Structural health monitoring in aviation: A comprehensive review and future directions for machine learning." *Nondestr.Test. Eval.* 40 (1): 1–60. https://doi.org/10.1080/10589759.2024.2350575.

Kulkarni, N. N., and A. Sabato. 2024. "Full-field expansion and damage detection from sparse measurements using physics-informed variational autoencoders." *Struct. Health Monit.* https://doi.org/10.1177/14759217241289575.

Kumar, P., A. Sharma, and S. R. Kota. 2021. "Automatic multiclass instance segmentation of concrete damage using deep learning model." *IEEE Access* 9: 90330–90345. https://doi.org/10.1109/ACCESS.2021.3090961.

Liao, X., Q. Yan, L. Su, Y. Qiu, J. Ren, and C. Zhang. 2024. "Automatic assessment of freeze-thaw damage in concrete structures using piezoelectric-based active sensing approach and deep learning technique." *Eng. Struct.* 302: 117453. https://doi.org/10.1016/j.engstruct.2024.117453.

Lin, T.-Y., P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. 2017. "Feature pyramid networks for object detection." In *Proc., IEEE Conf. on Computer Vision and Pattern Recognition*, 2117–2125. Piscataway, NJ: IEEE.

Liu, Y., M. Hajj, and Y. Bao. 2022. "Review of robot-based damage assessment for offshore wind turbines." *Renewable Sustainable Energy Rev.* 158: 112187. https://doi.org/10.1016/j.rser.2022.112187.

Liu, Y., X. Tan, and Y. Bao. 2024. "Machine learning-assisted intelligent interpretation of distributed fiber optic sensor data for automated monitoring of pipeline corrosion." *Measurement* 226: 114190. https://doi.org/10.1016/j.measurement.2024.114190.

Luo, K., X. Kong, J. Zhang, J. Hu, J. Li, and H. Tang. 2023. "Computer vision-based bridge inspection and monitoring: A review." *Sensors* 23 (18): 7863. https://doi.org/10.3390/s23187863.

Mohamed, A. G., M. R. Abdallah, and M. Marzouk. 2020. "BIM and semantic web-based maintenance information for existing buildings." *Autom. Constr.* 116: 103209. https://doi.org/10.1016/j.autcon.2020.103209.

Mundt, M., S. Majumder, S. Murali, P. Panetsos, and V. Ramesh. 2019. "Meta-learning convolutional neural architectures for multi-target concrete defect classification with the concrete defect bridge image dataset." In *Proc., IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 11196–11205. Piscataway, NJ: IEEE.

Pan, Y., G. Zhang, and L. Zhang. 2020. "A spatial-channel hierarchical deep learning network for pixel-level automated crack detection." *Autom. Constr.* 119: 103357. https://doi.org/10.1016/j.autcon.2020.103357.

Ronneberger, O., P. Fischer, and T. Brox. 2015. "U-Net: Convolutional networks for biomedical image segmentation." In *Proc., Medical Image Computing and Computer-Assisted Intervention*, edited by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, 234–241. Cham, Switzerland: Springer.

Sabato, A., S. Dabetwar, N. N. Kulkarni, and G. Fortino. 2023. "Noncontact sensing techniques for AI-aided structural health monitoring: A systematic review." *IEEE Sens. J.* 23 (5): 4672–4684. https://doi.org/10.1109/JSEN.2023.3240092.

Sabouni, A. R. 2023. "Advances in reinforced concrete integrity and failure." In *Advances in Structural Integrity and Failure*. London, UK: IntechOpen.

Shi, Y., L. Cui, Z. Qi, F. Meng, and Z. Chen. 2016. "Automatic road crack detection using random structured forests." *IEEE Trans. Intell. Transp. Syst.* 17 (12): 3434–3445. https://doi.org/10.1109/TITS.2016.2552248.

Tan, X., J. Du, Q. Zhang, W. Meng, and Y. Bao. 2024. "Monitoring restrained shrinkage and cracks of ultra-high-performance concrete (UHPC) using distributed fiber optic sensors." *Constr. Build. Mater.* 422: 135789. https://doi.org/10.1016/j.conbuildmat.2024.135789.

Tan, X., P. Guo, X. Zou, and Y. Bao. 2022. "Buckling detection and shape reconstruction using strain distributions measured from a distributed fiber optic sensor." *Measurement* 200: 111625. https://doi.org/10.1016/j.measurement.2022.111625.

Tian, Y., G. Zhang, H. Ye, Q. Zeng, Z. Zhang, Z. Tian, X. Jin, N. Jin, Z. Chen, and J. Wang. 2023. "Corrosion of steel rebar in concrete induced by chloride ions under natural environments." *Constr. Build. Mater.* 369: 130504. https://doi.org/10.1016/j.conbuildmat.2023.130504.

Wang, F., J. Huang, and Y. Fu. 2025. "Convolutional neural network-based multimodal image information fusion for moisture damage assessment of cultural heritage buildings." *Measurement* 242: 115972. https://doi.org/10.1016/j.measurement.2024.115972.

Xiang, C., W. Wang, L. Deng, P. Shi, and X. Kong. 2022. "Crack detection algorithm for concrete structures based on super-resolution reconstruction and segmentation network." *Autom. Constr.* 140: 104346. https://doi.org/10.1016/j.autcon.2022.104346.

Xie, E., W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. 2021. "SegFormer: Simple and efficient design for semantic segmentation with transformers." *Adv. Neural Inf. Process. Syst.* 21: 12077–12090.

Yang, F., L. Zhang, S. Yu, D. Prokhorov, X. Mei, and H. Ling. 2020. "Feature pyramid and hierarchical boosting network for pavement crack detection." *IEEE Trans. Intell. Transp. Syst.* 21 (4): 1525–1535. https://doi.org/10.1109/TITS.2019.2910595.

Yasmin, T., D. La, K. La, M. T. Nguyen, and H. M. La. 2024. "Concrete spalling detection system based on semantic segmentation using deep architectures." *Comput. Struct.* 300: 107398. https://doi.org/10.1016/j.compstruc.2024.107398.

Yu, D., Z. He, and L. Ma. 2023. "A modified multiscale semantic segmentation network accounting for multi-level seismic damage features of PC structure." *J. Build. Eng.* 78: 107600. https://doi.org/10.1016/j.jobe.2023.107600.

Zhai, G., Y. Xu, and B. F. Spencer. 2025. "Bidirectional graphics-based digital twin framework for quantifying seismic damage of structures using deep learning networks." *Struct. Health Monit.* 24 (1): 86–110. https://doi.org/10.1177/14759217241231299.

Zou, Q., Y. Cao, Q. Li, Q. Mao, and S. Wang. 2012. "CrackTree: Automatic crack detection from pavement images." *Pattern Recognit. Lett.* 33 (3): 227–238. https://doi.org/10.1016/j.patrec.2011.11.004.