MethodsX 13 (2024) 102796

Contents lists available at ScienceDirect

MethodsX

journal homepage: www.elsevier.com/locate/methodsx

Automated crack localization for road safety using contextual u-net with spatial-channel feature integration ☆

Priti S. Chakurkar [a,b], Deepali Vora [a,*], Shruti Patil [c], Ketan Kotecha [a]

[a] Computer Science and Engineering, Symbiosis Institute of Technology Pune, Symbiosis International (Deemed University) (SIU), Lavale, Pune, Maharashtra, India
[b] School of Computer Engineering, Dr.Vishwanath Karad MIT WORLD, PEACE UNIVERSITY, Kothrud, Pune, Maharashtra, India
[c] Artificial Intelligence and Machine Learning (AIML) Department, Symbiosis Institute of Technology, Symbiosis International (Deemed University) (SIU), Lavale, Pune, Maharashtra, India

ARTICLE INFO

Method name:
Contextual U-Net with Spatial-Channel Feature Integration for Crack Localization

Keywords:
Improving road safety
Pixel-level crack segmentation
EfficientNet encoder
Hierarchical attention networks

ABSTRACT

Accurate and timely crack localization is crucial for road safety and maintenance, but image processing and hand-crafted feature engineering methods, often fail to distinguish cracks from background noise under diverse lighting and surface conditions. This paper proposes a framework utilizing contextual U-Net deep learning model to automatically localize cracks in road images. The framework design considers crack localization as a task of pixel-level segmenting, and analyzing each pixel in a road image to determine if it belongs to a crack or not. The proposed U-Net model uses a robust EfficientNet encoder to capture crucial details (spatial features) and overall patterns (channel-wise features) within the road image. This balanced approach helps the model learn effectively from both individual elements and the context of the images, leading to improved crack detection. A customized hierarchical attention mechanism is designed to make U-Net model contextually adaptive to focus on relevant features at different scales and resolutions for accurately localizing road cracks that can vary widely in size and shape. The model's effectiveness is demonstrated through extensive evaluations on the benchmarked and custom-made datasets.

Specifications table

Subject Area:	Engineering
More specific subject area:	*Automated Road Crack Detection and Analysis*
Protocol name:	*Contextual U-Net with Spatial-Channel Feature Integration for Crack Localization*
Reagents/tools:	*PyTorch*
Experimental design:	*In order to make the U-Net model contextually adaptive and focus on pertinent features at various scales and resolutions for precisely localising road cracks, which can vary greatly in size and shape, a customised hierarchical attention mechanism is devised. Comprehensive tests on the benchmarked and custom datasets show the usefulness of the model.*
Trial registration:	*Not applicable*
Ethics:	*Not applicable*

(*continued on next page*)

☆ Direct Submission *No any Co-submissions papers that have been submitted alongside an original research paper accepted for publication by another Elsevier journal.*

* Corresponding author.
E-mail addresses: priti.chakurkar@mitwpu.edu.in (P.S. Chakurkar), deepali.vora@sitpune.edu.in (D. Vora).

https://doi.org/10.1016/j.mex.2024.102796
Received 14 March 2024; Accepted 8 June 2024
Available online 20 June 2024

Value of the Protocol:

- *Precise localization of road crack is challenged by environmental variations and complex surface conditions, achieving reduce false rates is essential, as inaccuracies can impact road maintenance decisions.*
- *An automated framework is proposed using a contextual U-Net deep learning model integrating EfficientNet with a customized hierarchical attention for accurate crack localization in road images*
- *A strategic data enhancement method adopted during the training of UNet further enhanced feature generalization capability of model.*

Background

The early approaches to road crack detection were predominantly manual, depending heavily on visual inspections conducted by human inspectors. This method involved personnel systematically surveying roads and highways, often aided by basic photographic documentation, where inspectors visually assess the road's surface, identifying and documenting areas with visible cracks. This method, while direct but limited to only the observer's experience and expertise. The reliance on manual inspection methods was reported by [1-3], who outlined the time-consuming and labor-intensive nature of these practices. They noted that such methods, while being the industry standard, were prone to human error, inconsistency and limited in precision. In response to these challenges, various conventional detection approaches have been developed over the years. These methods have primarily focused on using image processing techniques, which involve capturing road surface images and analyzing them for crack patterns. Early methods in this domain relied on techniques such as thresholding, edge detection, and morphological operations. While effective to a certain extent, these methods were limited in their ability to handle complex imaging conditions, such as varying lighting, shadows, and background textures. The challenge was further compounded by the low contrast of cracks against the road surface and the presence of other markings or debris. As the limitations of traditional image processing techniques became apparent, the focus shifted towards more sophisticated methods. This shift saw the advent of machine learning in road crack detection, offering a more robust and adaptive approach. Machine learning algorithms, particularly in the realm of deep learning, have demonstrated a remarkable ability to learn and identify patterns in data, making them well-suited for the task of crack detection. The application of convolutional neural networks (CNNs), for instance, has significantly improved the accuracy of crack detection. These networks can automatically learn hierarchical feature representations from the data, reducing the need for manual feature engineering and allowing for better generalization across different road conditions. However, despite the advancements brought by machine learning, several limitations persist. One of the key challenges is the models ability to generalize across various environments and Overfitting on specific datasets.

Many existing solutions are trained on specific datasets and may not perform well when exposed to different road types or environmental conditions. Additionally, the computational complexity of deep learning models can be a barrier, especially when deploying these systems in real-time scenarios or on edge devices [4]. Therefore, with the increasing complexity of road networks and the variability of road conditions there is a need for an efficient, and robust system for automated road crack detection. This research introduces a sophisticated learning framework to precisely differentiate between cracked and intact road surfaces at the pixel level.

Description of protocol

Framework overview

The real-world road image often subjected to different variety of environmental factors that alter the appearance of cracks, making their detection more difficult. For example, different lighting conditions and weather impacts, such as rain or shadows from trees and buildings, can either conceal cracks or create misleading features that resemble cracks. As it is believed by data scientist, that adequate data quality alone holds 70% for the performance improvement, while 30% depends on the neural network model. Therefore, the proposed framework adopts effective preprocessing operations to enhance the robustness and generalization ability of the model during the training phase. This includes dataset preparation, inconstancies handling and augmentation that introduces random pixel variations, mimicking the effect of wreckage, shadows, and other common road anomalies. This technique is particularly effective in teaching the model to distinguish between actual cracks and superficial irregularities that could otherwise lead to false positives.

The framework further leverages an advanced UNet segmentation model with a customized attention mechanism, particularly a hierarchical attention layers introduced as a substitution of skip connection in the UNet network architecture. To be more specific, UNet is a deep convolutional segmentation network model which is popular for its symmetry between down-sampling and up-sampling paths. This architecture ensures detailed feature extraction while maintaining spatial continuity, enabling accurate localization and context-aware segmentation. In our framework, the UNet model is advanced by integrating an EfficientNet encoder, distinguished by its depth-wise separable convolutions. This integration allows for distinct learning of spatial and channel-wise features, thereby enriching the feature set for a more detailed, low-level, and precise detection of road cracks. The approach not only improves the granularity of segmentation but also ensures a robust representation of the intricate characteristics of road surfaces.

To enhance the model's capability for spatial and channel-wise feature discrimination, we introduce a hierarchical-attention mechanism. A hierarchical attention mechanism is custom-made attention layer introduced in the proposed UNet learning network to replace skip connections with two parallel modules, including a self-attention module and a multi-scale attention module. The self-attention mechanism focuses on identifying dependencies within the feature maps, enabling the model to highlight salient features over larger image regions. On the other hand, the multi-scale attention mechanism operates at various scales, ensuring that the model remains sensitive to cracks of all sizes, from the minutest hairline fractures to larger, more visible cracks. First, the self-attention layer

operates at a single scale but considers the entire image, enabling the model to understand global dependencies within the features. It assigns significance to each pixel relative to others across the entire spatial domain of the input image, regardless of their proximity. Basically, this module highlighting features that contribute most significantly to the presence of road cracks, thereby enhancing the model's sensitivity to pertinent spatial cues. Secondly, the multi-scale attention module fuses multi-scale features in the skip path using a self-attention mechanism. This mechanism analyzes both spatial and channel information, allowing the model to capture long-range dependencies and preserve important details across the entire image. The deployment of attention strategy in hierarchical manner not only refines the segmentation process but also augments the model's ability to adapt to the intricate textures and patterns present in road surface images, leading to improved detection and localization of road cracks.

Contextual UNet learning model

The proposed system considers input images from the benchmarked datasets that includes a wide variety of real-world road images illustrating different background conditions and crack types. In the field of road crack detection, datasets typically comprise pairs of color input images and their corresponding ground truths, which are binarized representations manually annotated by domain experts. However, these ground truths are subject to human error and may exhibit inconsistencies in the distribution between the input images and their labeled counterparts. Inconsistencies or imbalances between these pairs can lead to skewed training and suboptimal model performance. Recognizing this challenge, the framework adopts an algorithmic solution designed to mitigate such discrepancies, ensuring a balanced and reliable dataset for training our model. The proposed algorithm proactively identifies and rectifies such imbalances by enforcing a strict one-to-one correspondence between input images and their ground truths, thereby leading to a more robust and reliable training environment.

Algorithm: Dataset Balancing Procedure

Input: I_{gt} (Set of ground truth images), I (Set of input images)
Output: A synchronized dataset with each input image $i \in I$ having a corresponding ground truth image $gt \in I_{gt}$.
Procedure

1. Initialize two lists:
   X: List to hold filenames of ground truth images with file extensions removed.
   L: List to hold filenames of input images.
2. Populate list X by processing I_{gt}:
   $X \leftarrow \{x | x = \text{filename}(gt) - \text{ext}, \forall gt \in I_{gt}\}$
3. Populate list L by processing I:
   $L \leftarrow \{l | l = \text{filename}(i), \forall\ i \in I\}$
4. For each input image i in I, perform the following steps:
   Let X_i be the corresponding ground truth image filename derived by removing the extension from i.
   Check if X_i exists in list X:
   If X_i is not in X, mark i as 'Absent'.
5. Create a list A to record 'Absent' filenames.
6. Iterate over the 'Absent' list A, and for each 'Absent' filename i, perform the deletion:
   Remove the corresponding ground truth image $I_{gt}(X_i)$ from the ground truth set I_{gt}.
7. Perform a similar check for each ground truth image gt in I_{gt} to ensure that each gt has a corresponding input image in I.
8. The resulting sets I and I_{gt} should now be synchronized, with each $i \in I$ having a corresponding $gt \in I_{gt}$.

End of procedure

The input images of preprocessed dataset are then resized to 128*128*3. Under the preprocessing the proposed system also executed data augmentation operation, artificially expanding the dataset by creating modified versions of the input images through techniques such as rotation, scaling, and adding noise. The aim is to make the model robust against various real-world variations it will encounter. After preprocessing, the images are passed through the proposed contextual U-Net model as shown below in Fig. 1.

Fig. 1 shows the overall framework of the proposed contextual UNet model which includes three parts: encoder part, hierarchical attention mechanism, and decoder module. The encoder operates through a sequence of convolutional stages, each precisely calibrated to downscale spatial dimensions while concurrently amplifying the depth of feature channels. The encoder initiates with layers specific to detecting low-level features, such as edges and textures, which are essential for early crack identification. As the input progresses through the model, deeper layers abstract increasingly complex feature representations.

- *Efficient Spatial Reduction: This process is characterized by logically diminishing the spatial dimensions of the input image while accurately preserving essential features, executed through* max*-pooling and strided convolutions such that:* $S_{I+1} = P(S_I)$ where, S_{I+1} *denotes the spatial dimensions of the feature map at layer l+1, and P represents the pooling operation applied to the spatial dimensions Sl of the previous layer.*
- *Layer-wise Feature Progression: The encoder processes the input image of dimension 128×128×3 through a series of convolutional layers, each layer designed to further abstract and encode the semantic information present in the images. The progression through these layers basically performs scaling down the resolution and increasing depth highlighted as follows:*
  - *Input Resolution: The initial input resolution of the images is 128×128×3. This 3 represents the three-color channels (RGB) in the image. The spatial dimensions (128×128) indicate the height and width of the image in pixels.*
  - *First Convolutional Layer: The input is first processed through a convolutional layer that reduces the spatial dimensions to 64×64, while depth (number of feature channels) increases to 64. This layer focuses on extracting primary features such as edges and basic textures.*

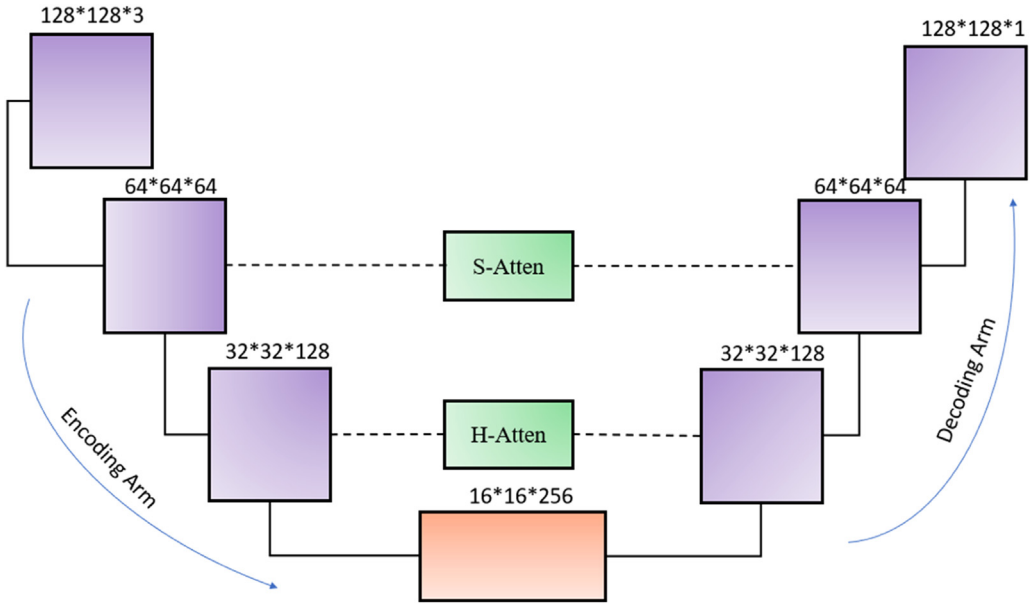

Fig. 1. Proposed contextual U-net with spatial-channel feature integration.

This transformation can be regarded as a transformation from a higher resolution with lower feature depth to a lower resolution with higher depth.

- *Second Convolutional Layer: The subsequent layer further downscales the spatial dimensions to 32×32, and the depth of the feature maps is increased to 128. This increase in depth allows the network to encode more complex features and patterns, which are essential for differentiating cracks from other road surface anomalies.*
- *Third Convolutional Layer: The final convolutional layer in the sequence reduces the spatial dimensions to 16×16, with the feature depth further increasing to 256. This layer is significant in capturing the most abstract and semantic information from the input, which includes understanding the finer details and variations in the road surface cracks.*

The use of progressively deeper convolutional layers is a strategic approach in the Efficient U-Net Encoder. As the network reaches deeper, it becomes increasingly capable of encoding richer contextual information. In our enhanced UNet framework, we incorporate a hierarchical attention mechanism to significantly improve the model's ability in distinguishing between crack and non-crack regions at the pixel level. This mechanism, replacing traditional skip connections, consists of two parallel modules: a self-attention module and a multi-scale attention module, uniquely designed to capture and emphasize crucial features for precise road crack detection.

- The self-attention module is strategically positioned after initial convolutional layers to refine feature maps by capturing global dependencies, thereby focusing the network's attention on regions most likely to contain cracks.
- The multi-scale attention module, placed deeper within the encoder, amplifies features extracted at various scales. This ensures the model's attentiveness to significant features for crack detection, which can vary in size, shape, and texture.
- Both attention module works to refine the encoder's output before it is passed to the decoder. The self-attention module computes attention scores to reweight the features, making the model more adept at distinguishing cracks from complex road surface backgrounds. The implementation involves convolutional operations to generate Q, K, and V from the input feature map, followed by the matrix multiplication and SoftMax operations to compute the attention scores and the subsequent reweighting of features.

Fig. 2 shows a schematic representation of simple multi-scale attention workflow. The process starts with the extraction of features at multiple scales via pyramid pooling, which is designed to capture a broader context as the filter size increases. This step ensures that the network perceives the input image at different resolutions, making it sensitive to features of varying dimensions.

Here, each scale independently undergoes an attention process, similar to self-attention, where an attention matrix is used to calculate attention weights. These weights determine the importance of features at each particular scale, allowing the network to focus more on significant features and neglect less useful information. In the Latter stages the features from each scale are integrated to form a comprehensive feature map.

Impact on Road Crack Segmentation:

- The hierarchical attention mechanism's selective focus and ability to handle varied textures lead to an improvement in segmentation accuracy. It enables the model to selectively concentrate on pertinent areas within road images that are more likely to contain cracks, enhancing segmentation precision.

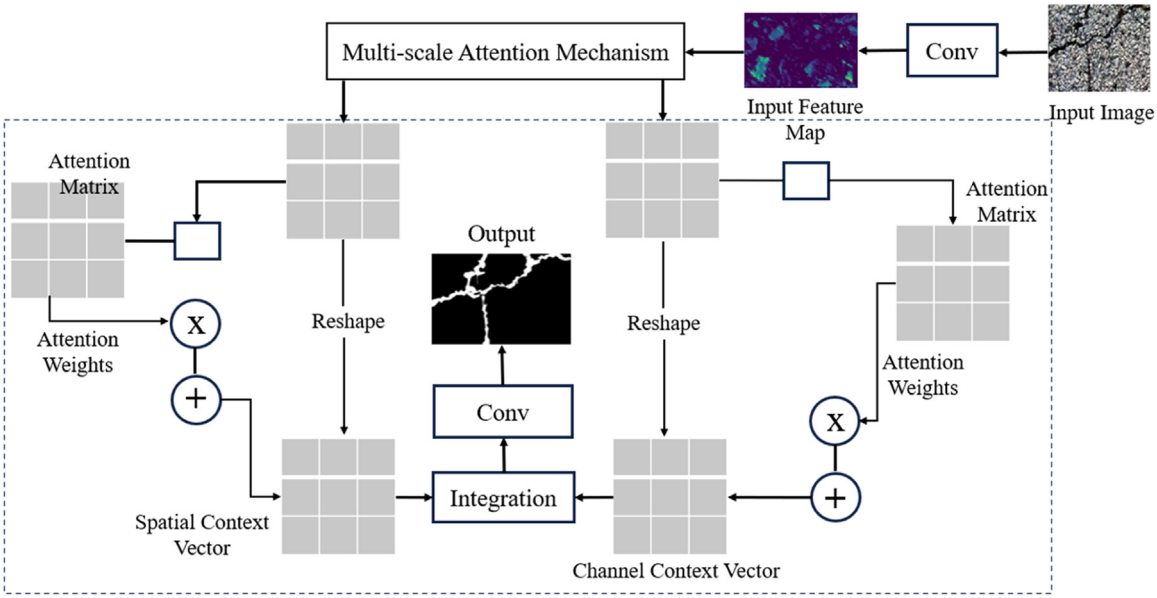

Fig. 2. Schematic workflow of multiscale attention mechanism.

Table 1
Summary of the proposed UNet model.

Module Type	Input Shape	Convolution Kernel	Output Shape	Repeated Times	Stride
Initial Convolution	128×128×3	3×3	64×64×64	1	2
Convolution + Pooling	64×64×64	3×3	32×32×128	1	2
Convolution + Pooling	32×32×128	3×3	16×16×256	1	2
Self-Attention	16×16×256	-	16×16×256	1	-
Multi-Scale Attention	Various	-	Match encoder outputs	1 per skip path	-
Up-sampling + Convolution	16×16×256	3×3	32×32×128	1	-
Concatenation + Convolution	32×32×128 + 32×32×128	3×3	32×32×128	1	-
Up-sampling + Convolution	32×32×128	3×3	64×64×64	1	-
Concatenation + Convolution	64×64×64 + 64×64×64	3×3	64×64×64	1	-
Up-sampling + Convolution	64×64×64	3×3	128×128×32	1	-
Final Convolution	128×128×32	1×1	128×128×1	1	-

- By integrating attention mechanisms in a hierarchical manner, our model adapts to complex road textures and patterns, significantly reducing false positives improving the detection and localization of road cracks. This not only refines the segmentation process but also strengthens the model's overall performance in road crack detection tasks.

In our enhanced context-aware U-Net architecture, the decoder module reconstructs the segmented image from encoded features, effectively utilizing insights from the hierarchical attention layer. The key function of the decoder module is discussed as follows:

- **Up-sampling Process**: The decoder progressively increases the spatial resolution of feature maps received from the encoder. This process is mathematically defined *as* $U_{l+1} = Up(U_l)$, where U_{l+1} is the up-sampled feature map and Up signifies the up-sampling operation.
- **Integration of Attention-Enhanced Features**: Unlike traditional skip connections, our model incorporates features refined by the attention mechanisms. This ensures that the decoder focuses on the most relevant features for crack segmentation, enhancing model accuracy and adaptability to complex road textures.
- **Feature Refinement**: Following the feature integration, the decoder conducts convolutional operations to refine the combined features further, defined as $R_l = Conv(D_l; W_l)$, where R_l represents the refined feature map. This step is crucial for preserving critical details, such as small cracks, throughout the up-sampling process.

The decoder module is implemented as a series of convolutional layers followed by up-sampling operations, is designed to reintegrate attention-focused features back into the segmentation map. This process not only counteracts the limitations of conventional U-Net architectures but also leverages the hierarchical attention mechanisms to adaptively highlight key features, thereby significantly enhancing the segmentation outcomes for road crack detection. Table 1 summarizes the architecture of the proposed contextual U-Net architecture.

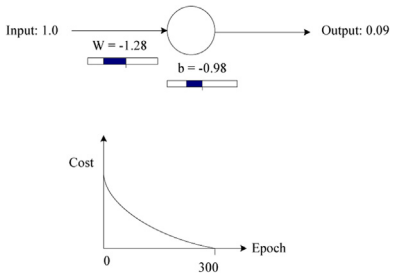

Fig. 3. Cost Vs Epoch.

The design and development of the proposed advanced U-Net model for precise road crack localization is carried out using Python programming language within an Anaconda distribution. The training process considers both training and validation set, with Stochastic gradient descent (SGD) optimizer and a hybrid loss function combining Dice score and binary cross-entropy (BCE) loss function. The Dice score, a widely recognized loss function for segmentation tasks, was utilized to optimize the model. This choice is motivated by the Dice score's effectiveness in handling class imbalance, a common challenge in pixel-level segmentation tasks such as road crack detection. The Dice loss among two binary vortexes is as in the Eq. (1).

$$D_L = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2} \tag{1}$$

Where the sums run over the N voxels, of the predicted binary segmentation volume $g_i \in P$ and the ground truth binary volume $g_i \in G$. This formulation of Dice can be differentiated w concerning the $j-th$ voxel of the prediction, yielding the gradient: as in Eq. 2.

$$\frac{\partial D_L}{\partial p_i} = 2\left[\frac{g_i\left(\sum_i^N p_i^2 + \sum_i^N g_i^2\right) - 2p_j\left(\sum_i^N p_i g_i\right)}{\left(\sum_i^N p_i^2 + \sum_i^N g_i^2\right)}\right] \tag{2}$$

While in theory, DICE loss seems to be an ideal measure for the error in the output, practically speaking, it is not very good. The reason is that the output pixel value is not just 0 or 1 it is a value between 0 and 1. While measuring the dice loss, the output value of each pixel is rounded to either 0 or 1. The main issue with this is say the output value is 0.501, then it will be rounded to 1.0. however, this is not correct. Dice loss does indeed measure the overall correctness of the image. However, it does not give us the correctness of each pixel. To avoid this, both BCE loss and DICE loss are considered. In the case of BCE loss, if the expected value is opposite to the predicted value, then the loss is infinite and it gives a value between 0 and infinity to all other values. This BCE loss value is measured for each pixel scaled down to a value between 0 and 1 and then averaged. Thereby, we will ensure not only the overall image correctness but pixel-wise correctness as well. In line with the different activation function, the binary cross entropy has been considered due to its circumstantial requirements.

For the same given set of input-output pair and activation function Sigma the mean square error, with the parameters that include, desired output y = 0.0, for corresponding input x = 1.0 with bstart = 0.9, wstart =0.6 to satisfy the condition z=wx+b the objective function is $a = \sigma(z)$, the cost function is as in Eq.(3)

$$C = \frac{(y-a)^2}{2} \tag{3}$$

The pattern of the cost value with respect to each epoch is as in the Fig. 3

However, for same set of y and x, if the values of bias and weight is $b_{start} = 2.0$, $w_{start} = 2.0$ then for the Eq. (4)

$$\frac{\partial C}{\partial w} = (a-y)\sigma'(z)x = a\sigma'(z) \tag{4}$$

However, if the cost function is as in Eq. (6),

$$C = -\frac{1}{n}\sum[y\ In\ a + (1-y)In\ (1-a)] \tag{6}$$

Which further generalizes as

$$\frac{\partial C}{\partial w_j} = -\frac{1}{n}\sum_x\left(\frac{y}{\sigma(z)} - \frac{(1-y)}{1-\sigma(z)}\right)\frac{\partial \sigma}{\partial w_j} = -\frac{1}{n}\sum_x\left(\frac{y}{\sigma(z)} - \frac{(1-y)}{1-\sigma(z)}\right)\sigma'(z)x_j.$$

$$\frac{\partial C}{\partial w_j} = \frac{1}{n}\sum_x \frac{\sigma'(z)x_j}{\sigma(z)(1-\sigma(z))}(\sigma(z)-y).$$

$$\frac{\partial C}{\partial w_j} = \frac{1}{n}\sum_x x_j(\sigma(z)-y)$$

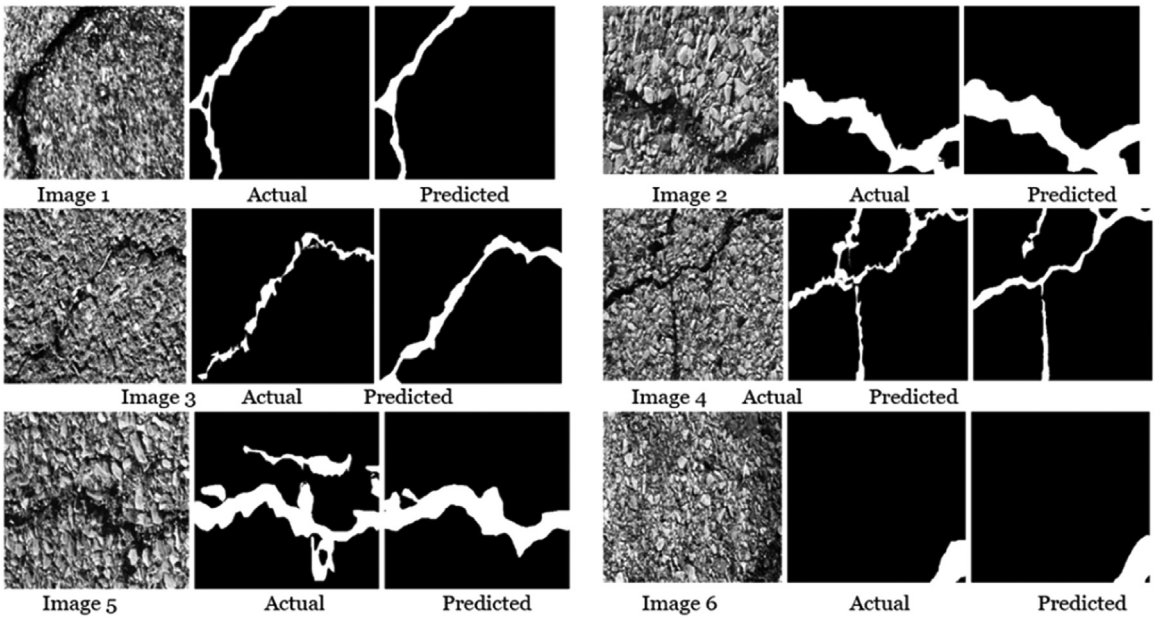

Fig. 4. Visual analysis of input, ground truth and segmented output image.

For any value of weight and bias for given pair of input-desired output, the curve remain same. This is the cost function is used in the design of the road crack location learning model, where the binary classification takes place. The model is trained for 50 epochs and the evaluation of model performance is carried out through standard segmentation metrics such as mean accuracy, precision, recall, F1-score and Mean Intersection over Union (MIU/IoU) that evaluates the overlap between predicted and actual crack segments, serving as a critical indicator of segmentation success. Extensive analysis is carried out considering two benchmarked and custom-made datasets. The following results are shown for benchmarked dataset Crack500 are the result are compared with similar existing research works.

Fig. 4 shows the visual outcomes of the proposed contextual U-Net model through a series of side-by-side comparisons between actual and predicted segmentation of road cracks. Each test instance row presents a distinct road surface image (Image 1 to Image 6), with ground truth (Actual) and the model's prediction (Predicted) for crack localization. Based on the closer analysis it can be seen that the model accurately outlines the crack, indicating model's capability to identify and follow the intricate pathways of the cracks, which is particularly challenging given the irregular patterns and varying widths. These visual outcomes not only validate the efficacy of the proposed model in accurate crack segmentation but also emphasize its potential application in real-world scenarios where such precision is crucial for timely road maintenance and safety measures.

Fig. 5 demonstrates model performance in terms of different statistical measures. Fig. 5(a) depicts the analysis of the Dice coefficient for both training and validation phases over a series of epochs. The Dice coefficient is a measure of the model's segmentation performance, with a value ranging from 0 to 1, where 1 signifies perfect and complete overlap between the predicted and ground-truth segments. The training Dice score demonstrates a steady and consistent improvement, maintaining high values throughout the epochs, which indicates that the model is effectively learning the crack segmentation task. The validation Dice score remains stable after an initial period of improvement, which suggests that the model has achieved a robust generalization capability and is performing well on unseen data without overfitting.

Fig. 5(b) shows performance analysis regarding training and validation IoU. It can be seen that the training Intersection IoU shows a similar trend to the Dice score, with a high and consistent performance, reflecting the model's accuracy in predicting crack pixels as the training progresses. Validation IoU, though slightly more variable, stays within a relatively tight range, indicating the model's dependable performance on the validation set and its ability to capture the true positive rate effectively.

Fig. 5(c) shows performance analysis regarding training and validation precision. The precision on the training set is quite high, suggesting that when the model predicts the presence of a crack, it is correct most of the time, which is crucial for reducing false positives in crack detection. The validation precision exhibits some fluctuation but remains within an acceptable range. The higher variability could be due to the model encountering a variety of crack presentations in the validation set, yet it still demonstrates the model's strong predictive precision.

Fig. 5(d) shows performance analysis concerning training and validation precision. It can be also seen that the training recall is consistently high, which means that the model is successfully identifying a large proportion of actual crack pixels during the training phase. The validation recall, while demonstrating greater variability, indicates the model's sensitivity to detecting cracks, ensuring that fewer actual cracks go undetected. Based on the overall analysis it can be observed that the proposed model learns well and generalizes effectively to new data. The high training scores across all metrics indicate a strong learning capability, and the stability

Fig. 5. Analysis of Model performance.

Table 2
Comparative analysis.

Models	*Precision*	*Recall*	*F1*	*IOU/MIU*	*Accuracy*
Res-UNet [5]	*74.26%*	*72.85%*	*73.27%*	-	-
CTN [6]	*69.1%*	*78.0%*	*73.3%*	-	-
U-Net [7]	*85.34*	*68.13*	*75.7*	*62.48*	-
RUC-Net [8]	*69.88%*	*76.19%*	*72.90%*	*57.36%*	-
DeepCNN [9]	*65.4%*	*69.8%*	*67.5%*	*73.5%*	*83.8%*
Proposed	*75.81%*	*78.11%*	*75.88%*	*74.93*	*97.49*

of the validation scores, despite some variability, confirms that the model is not overfitting and can maintain its performance on unseen data.

The above Fig. 6 presents additional outcome in response to the concern about overfitting as indicated by the results in Fig. 5. Overfitting often arises when a model learns the details and noise in the training data to an extent that it negatively impacts the performance on new data. Here, the demonstration of the model's performance on additional, complex real-world images not included in the initial training set showcases its ability to generalize beyond the controlled conditions of the training data. These results suggest that while the model may show high performance on the training and slight less performance on validation sets, it still retains a significant capacity to perform effectively on new, unseen data types. Therefore, to ensure the model's generalization capabilities, the performance on high-resolution and unsegmented images are shown in Fig. 6 where first image is from real-time captured images and remaining two are randomly selected from internet sources. The outcome produced by the model is promising with complex crack scenario, which not only validate the model's ability to handle diverse and complex inputs but also showcase its effectiveness in accurately segmenting cracks under less controlled and varied conditions. Originally, training the model on cropped images was essential to focus on specific features and improve learning efficiency. However, transitioning to testing on high-resolution and complex images demonstrates the model's adaptability and robustness.

Table 2 compares the proposed model against existing approaches (Res-UNet, CTN, U-Net, RUC-Net, DeepCNN) on the Crack500 dataset. The proposed model achieves superior performance across several key metrics. It can be observed that the proposed model

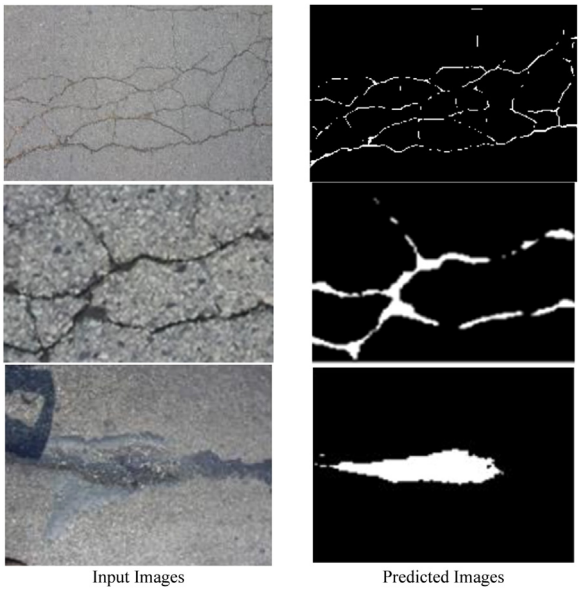

Fig. 6. Trained model performance analysis with complex crack images.

outperforms the others with a precision of 75.81%, indicating that it has a higher rate of true positive predictions. This suggests that the proposed model is less prone to false positives. At 78.11%, the recall of the proposed model is also superior, suggesting it is better at identifying all relevant instances of road cracks compared to the other models. The F1-score for the proposed model is 75.8%. The proposed model achieves an IoU of 74.9%, and accuracy of 97.49%, significantly higher than the other models. Hence Based on the overall analysis of the outcome, it can be seen that the proposed model demonstrates a significant improvement in precision, recall, and accuracy over existing models like Res-UNet, CTN, U-Net, RUC-Net, and DeepCNN. While the IoU is not much highest, it is still more than other models. The improved precision and recall suggest that the proposed model is effective in reducing both false positives and false negatives, making it a robust approach for road crack detection and potentially reducing the costs associated with road maintenance and repair.

In the comparative study presented in Table 2, fairness and rigor were maintained across the evaluation of different models. Each model, including the proposed one, underwent a standardized process for hyperparameter tuning to avoid biases and local optima. We employed a grid search strategy, ensuring that each model's hyperparameters were optimized based on a consistent criterion across all models, specifically targeting the maximization of the F1-score on a held-out validation set. Furthermore, to address potential issues of local optima and ensure robust generalization, all models were trained using the same initialization strategies and learning rate schedules, and training was extended until no significant improvement was observed on the validation set for a predefined number of epochs. The training and validation splits were identical for all models, and results were averaged over multiple runs to account for variability in training dynamics.

Limitations

The proposed model achieves an IoU of 74.9%, which, while not the highest, is competitive and suggests good overlap between the predicted segmentations and the ground truth. Although there are similar existing techniques are proposed in literature like work in [10] presented lightweight semantic segmentation for bridge inspection. Here, the authors have adapted DeepLabv3+ using MobileNetV2 as the backbone, employing depthwise separable and atrous convolutions to enhance efficiency, particularly suitable for deployment on edge devices. While this approach focuses on operational efficiency in edge computing scenarios, our contextual U-Net model is designed to balance efficiency with high accuracy across standard computing environments. Our model achieves a superior intersection over union (IoU) score, indicative of better segmentation quality without the need for hardware-specific optimizations. This makes it versatile and easily deployable across various standard devices without specialized architectural adjustments. Another work done by [11] Utilizes a modified U-Net architecture incorporating Self-Attention-Self-Adaption (SASA) neurons. The specialized SASA neurons and CRED algorithm focus on tiny, complex crack patterns. In contrast, our contextual U-Net effective in identifying and segmenting both complex and broader crack types without requiring such specific adaptations. This generality allows for wider application across different scenarios without the need for retraining or modifying the underlying neural structure. In [12] A task-aware meta-learning model is presented for structural damage aanalysis using Model-Agnostic Meta-Learning (MAML) to enhance generalization across diverse structural damages with limited image sets. However, our model does not currently employ meta-learning, its high performance on standard datasets and the ability to generalize well to new, unseen data still make it a competitive tool in crack detection. Integrating meta-learning could be a future step to broaden its applicability across a more diverse set of

structural damages. In the study of [13] a modified fusion cnn for crack identification is presented to handle complex disturbances, including multi-level and multi-scale features, using super-resolution processes to enhance identification accuracy. The propsoed framework does not specifically customized for super-resolution inputs, maintains high accuracy and robustness in standard-resolution environments, ensuring reliable performance without the computational overhead associated with super-resolution processing. This ensures that our model remains efficient while still providing precise crack detection, which is crucial for practical applications where processing speed and resource efficiency are priorities.

Ethics statements

This research did not involve human participants, animal experiments, or data collected from social media platforms. All road image data utilized in this study are sourced from publicly available datasets, which were collected and made available by researchers adhering to the respective ethical guidelines and without violating privacy rights. No additional ethical approval was required for the use of these datasets in our study.

CRediT author statement

Priti Chakurkar: Conceptualization, Methodology, Software, Data Curation, Writing - Original Draft, Validation, Visualization. Led the design and development of the automated crack localization framework, implemented the contextual U-Net learning model, prepared and analyzed the datasets, wrote the original draft of the manuscript, conducted validation tests, and generated visualizations for the study. **Deepali Vora:** Supervision, Writing - Review & Editing, Project Administration, Funding Acquisition. Provided overall project guidance, critically reviewed and edited the manuscript for intellectual content, administered the project's execution, and secured funding for the research. Shruti Patil: Provided overall project guidance, critically reviewed and edited the manuscript for intellectual content, administered the project's execution, and secured funding for the research. **Ketan Kotecha:** Provided overall project guidance, critically reviewed and edited the manuscript for intellectual content, administered the project's execution, and secured funding for the research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] H. Oliveira, P.L. Correia, Automatic Road cracks detection and characterization, IEEE Trans. Intell. Transp. Syst. 14 (2012) 155–168.
[2] U.A. Nnolim, Automated crack segmentation via saturation channel thresholding, area classification and fusion of modified level set segmentation with Canny edge detection, Heliyon 6 (12) (2020) e05748.
[3] S.Y. Alam, A. Loukili, F. Grondin, E. Rozière, Use of the digital image correlation and acoustic emission technique to study the effect of structural size on cracking of reinforced concrete, Eng. Fract. Mech. 143 (2015) 17–31.
[4] P.S. Chakurkar, D. Vora, S. Patil, S. Mishra, K. Kotecha, Data-driven approach for AI based crack detection: Techniques, challenges, and future scope, Front. Sustain. Cities 5 (2023) 1253627.
[5] S.L.H. Lau, E.K.P. Chong, X. Yang, X. Wang, Automated pavement crack segmentation using U-net-based convolutional neural network, IEEE Access 8 (2020) 114892–114899.
[6] H. Tao, B. Liu, J. Cui, and H. Zhang, A convolutional-transformer network for crack segmentation with boundary awareness, arXiv [cs.CV], (2023).
[7] D. Benedetto, A. Fiani, M. Gujski, U-Net-Based CNN Architecture for Road Crack Segmentation, Infrastructures 8 (5) (2023).
[8] G. Yu, J. Dong, Y. Wang, X. Zhou, RUC-Net: A residual-Unet-based convolutional neural network for pixel-level pavement crack segmentation, Sensors (Basel) 23 (1) (2022).
[9] Z. Qu, C. Cao, L. Liu, D.-Y. Zhou, A deeply supervised convolutional neural network for pavement crack detection with multiscale feature fusion, IEEE Trans. Neural Netw. Learn. Syst. 33 (9) (2022) 4890–4899.
[10] Xu, Yang, Fan, Yunlei, Qiao, Weidong & Li, Hui. (2022). Lightweight deep learning model of semantic segmentation for complex concrete cracks in actual bridge inspection. 10.12783/shm2021/36273.
[11] J. Zhao, F. Hu, W. Qiao, W. Zhai, Y. Xu, Y. Bao, H. Li, A modified U-net for crack segmentation by Self-Attention-Self-Adaption neuron and random elastic deformation, Smart Struct. Syst. 29 (1) (2022) 1–16.
[12] Y. Xu, Y. Fan, Y. Bao, H. Li, Task-aware meta-learning paradigm for universal structural damage segmentation using limited images, Eng. Struct. 284 (2023) 115917.
[13] Y. Xu, Y. Bao, J. Chen, W. Zuo, H. Li, Surface fatigue crack identification in steel box girder of bridges by a deep fusion convolutional neural network based on consumer-grade camera images, Struct. Health Monitoring 18 (3) (2019) 653–674.