Optimizing Crack Detection: The Integration of Coarse and Fine Networks in Image Segmentation

Hoanh Nguyen*, Tuan Anh Nguyen
Faculty of Electrical Engineering Technology, Industrial University of Ho Chi Minh City, Ho Chi Minh City, Vietnam

***Abstract*—In recent years, the automation of detecting structural deformities, particularly cracks, has become vital across a wide range of applications, spanning from infrastructure maintenance to quality assurance. While numerous methods, ranging from traditional image processing to advanced deep learning architectures, have been introduced for crack segmentation, reliable and precise segmentation remains challenging, especially when dealing with complex or low-resolution images. This paper introduces a novel method that adopts a dual-network model to optimize crack segmentation through a coarse-to-fine strategy. This model integrates both a coarse network, focusing on the global context of the entire image to identify probable crack areas, and a fine network that zooms in on these identified regions, processing them at higher resolutions to ensure detailed crack segmentation results. The foundation of this architecture lies in utilizing shared encoders throughout the networks, which highlights the extraction of uniform features, paired with the introduction of separate decoders for different segmentation levels. The efficiency of the proposed model is evaluated through experiments on two public datasets, highlighting its capability to deliver superior results in crack detection and segmentation.**

***Keywords*—Deep learning; crack segmentation; coarse-to-fine strategy; image segmentation**

I. Introduction

The structural integrity and safety of infrastructures, such as buildings, roads, dams, and bridges, are vital to the well-being of societies across the world. One of the earliest and most common indicators of deteriorating structural health is the appearance of cracks. Crack segmentation, an essential branch of computer vision and structural health monitoring, focuses on the accurate identification and tracking of cracks in various materials and surfaces. The primary objective is to detect cracks as early as possible, ensuring timely maintenance, prevention of potential catastrophic failures, and extension of the lifespan of structures. With the rapid development of deep learning [1], [2], [3], [4] and the emergence of segmentation models such as SegNet [5], UNet [6], FCNs [7], and the DeepLab series [8], [9], general semantic segmentation tasks and crack segmentation tasks, in particular, have achieved significant improvements [10], [11], [12]. However, crack segmentation still poses numerous challenges that need addressing [13], [14], [15]. First, cracks can appear in various shapes, ranging from fine lines to wide gaps, and may display in different depths, lengths, and orientations. This makes it challenging for models to generalize across all possible crack presentations. Second, the surface on which a crack appears often possesses its texture, which can resemble a crack, making it difficult to differentiate between actual defects and background patterns. Third, uneven and dynamic lighting conditions, as well as external factors such as dirt, moisture, or staining can either obscure cracks or create shadows that might be mistaken for cracks. In recent years, with the rapid advancement of convolutional neural networks (CNNs), many methods have been proposed to address these challenges. In study [16], the authors proposed a novel unsupervised multi-scale fusion crack detection algorithm for pavement images, which addresses challenges posed by intensity inhomogeneity, topology complexity, and other factors without the need for training data. This method integrates a windowed minimal intensity path-based technique for candidate crack extraction, cross-scale crack correspondence, and a multivariate statistical hypothesis test for crack evaluation. In study [17], the authors introduced a cutting-edge network architecture called feature pyramid and hierarchical boosting network tailored for pavement crack detection. This network integrates context information into low-level features through a feature pyramid approach, and introduces a unique nested sample reweighting process, along with a new measurement method, the average intersection over union for enhanced crack detection accuracy. Yue et al. [18] presented CrackNet-V, an enhanced deep network tailored for pixel-level crack detection in 3D asphalt pavement images. This advanced network, building upon the foundational principles of CrackNet, features a deeper structure with fewer parameters, thereby offering superior accuracy and computational efficiency, while also incorporating novel features like the leaky rectified tanh activation function for precise shallow crack detection. Zhengxin et al. [19] proposed a semantic segmentation neural network designed for road area extraction, seamlessly merging the capabilities of residual learning and the U-Net architecture. This model incorporates residual units for simplified deep network training while its rich skip connections streamline information propagation, resulting in a leaner yet more performative network. In study [20], the authors introduced the Crack Transformer network (CrackFormer), a specialized solution tailored for fine-grained crack detection, integrating innovative attention mechanisms within a SegNet-inspired encoder-decoder framework. CrackFormer features unique self-attention modules and efficient positional embedding, while also incorporating new scaling-attention modules, emphasizing semantic crack features and reducing non-semantic interferences.

Although the above methods have addressed many challenges of crack segmentation, some difficulties remain, especially with thin cracks. As illustrated in Fig. 1, thin cracks are often more difficult to detect, particularly in low-resolution images. Furthermore, thin cracks can easily be mistaken for the

natural texture of asphalt, especially when the asphalt surface is rough or granular. Thin cracks can also appear darker or fainter depending on the lighting conditions and the angle of image capture, posing challenges for consistent detection. To address these issues, this paper introduces a dual-network model that employs a coarse-to-fine strategy for enhanced crack segmentation. The model consists of two networks: the coarse network captures global image context to identify potential crack areas, followed by the fine network, which focuses on these identified regions at a high resolution to achieve precise segmentation. Both networks share an encoder, but use separate decoders to process images, ensuring high-quality crack detection results. The proposed model is evaluated on two public datasets, including CrackTree260 and DeepCrack537 datasets. Experimental results show that the proposed model delivers superior results in crack detection and segmentation.

Fig. 1. Some images illustrate the challenges faced when performing crack segmentation.

II. Proposed Model

Fig. 2 provides a detailed visualization of the multi-stage segmentation process implemented in the proposed method. The proposed pipeline consists of both a coarse and a fine network. While both networks use a shared encoder, denoted as E, they each have their own decoders: D_c for the coarse network and D_f for the fine network. The main objective of the coarse network is to capture global contextual information from the entire image and subsequently highlight regions potentially containing cracks. Based on the predictions from the coarse network, the fine network then zooms into these identified regions, focusing specifically on local patches considered to have cracks, to achieve high-resolution crack segmentation. An input image represented mathematically as $I \in R^{W \times H \times 3}$ undergoes a downsampling step first. This optimizes its size for an initial analysis without sacrificing key details. Following this, the coarse network processes the downsized image to identify regions that may have cracks, resulting in a coarse crack map. This map essentially serves as a probability output, given by $P = Sigmoid(D_c(E(I)))$, highlighting pixels with a higher probability of being part of a crack. To refine this output, a thresholding technique is applied, producing a more defined coarse crack mask. This mask is then employed as a guide for the subsequent fine network. The fine network delves deeper, cropping and zooming into the previously identified regions. Its primary task is to work on these local patches, operating at a higher resolution to accurately segment the cracks, leading to a precise segmentation result.

A. Shared Encoder

We use ResNet-101 architecture in [21] as the shared encoder of our model. ResNet-101, a variant of the Residual Network (ResNet) family, is a deep convolutional neural network architecture known for its excellent performance on a variety of computer vision tasks. At its core, the design philosophy behind ResNet-101 is the introduction of "residual blocks" which address the vanishing gradient problem encountered in very deep neural networks. The network begins with a single convolutional layer with a 7×7 kernel, stride of 2, followed by a max pooling layer. The majority of the network consists of sequences of residual blocks. These blocks allow the model to learn identity functions that ease the training of deeper networks by providing shortcut connections across layers. Specifically, a residual block contains a skip connection that bypasses one or more layers. ResNet-101 contains four main groups of residual blocks. The first group has 3 blocks, the second group has 4 blocks, the third group comprises a significant 23 blocks, and the fourth group has 3 blocks. Each block within these groups consists of 3 layers (i.e., a 1×1 convolution, a 3×3 convolution, and another 1×1 convolution), with the exception of the first block of each group, which adjusts the number of channels and downsamples using a stride of 2. For crack segmentation task, we remove fully connected layers at the end of the architecture to maintain spatial information throughout the network. In addition, to stabilize the activations and speed up training, batch normalization and ReLU (Rectified Linear Unit) activation functions are applied after each convolution within the blocks. The details of ResNet-101 structure used in this paper are shown in Table I. The deep nature of ResNet-101 enables it to capture a wide range of features at different scales, while its residual connections ensure that training remains stable and efficient even with its impressive depth.

B. Segmentation Decoders

For both coarse and fine segmentation decoders, this paper employs the Atrous Spatial Pyramid Pooling (ASPP) decoder in [22]. ASPP is a prominent module designed to capture multi-scale contextual information without the need for multiple input scales or exhaustive downsampling. Originally proposed for semantic image segmentation in the context of the DeepLab series of architectures, ASPP is specifically designed to handle the challenges posed by objects of varying scales in images. The core concept behind ASPP is to apply parallel dilated (or atrous) convolutions on the feature map, each having a different dilation rate. This results in capturing spatial information from different field-of-views without substantially increasing the number of parameters or the computational burden. The multi-scale feature maps resulting from these parallel operations are then concatenated. Specifically, the ASPP module comprises: A 1×1 convolution which captures the image's immediate context, three 3×3 convolutions but with varying dilation rates (e.g., 6, 12, and 18), which allow the network to capture spatial information from different ranges without downsampling, and a global average pooling layer to

process the feature map, capturing the holistic context of the image. The resulting features are then upsampled and concatenated with the other components. After concatenating the outputs from these parallel operations, the combined feature map passes through another convolution to produce the final enhanced feature map that fuses multi-scale contextual information. This feature map is then fed into the ASPP decoder for semantic segmentation, as shown in Fig. 3. In the ASPP decoder, the enhanced feature maps are first bilinearly upsampled by a factor of 4 and then concatenated with the corresponding low-level feature map from the backbone (i.e., Conv2 layer of ResNet-101). To implement the concatenation operation, a 1×1 convolution layer is applied to the low-level feature map to reduce the number of channels. After concatenation, two 3×3 convolution layers are used to refine the features, followed by another simple bilinear upsampling by a factor of 4. The refined feature map is finally fed into the segmentation head. The ASPP decoder enables the network to effectively segment objects of various sizes and scales in images, making it a powerful choice for many semantic segmentation tasks.

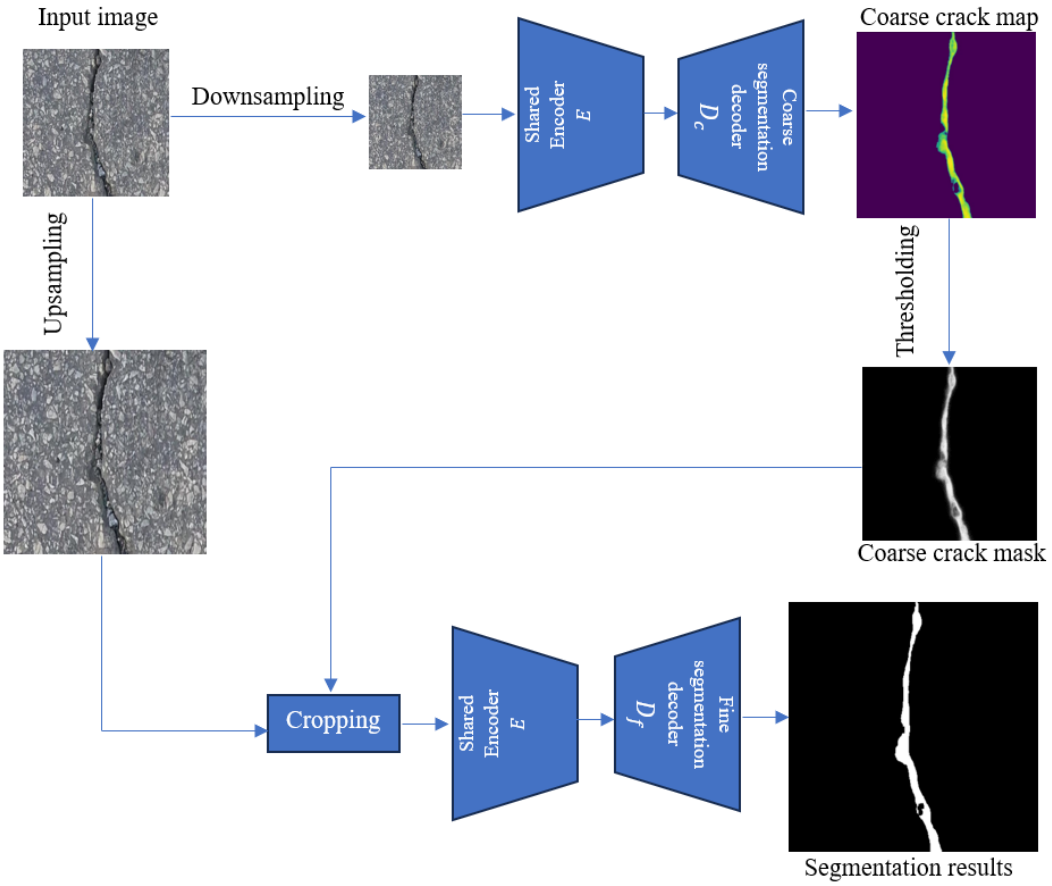

Fig. 2. The structure of the proposed model for crack segmentation.

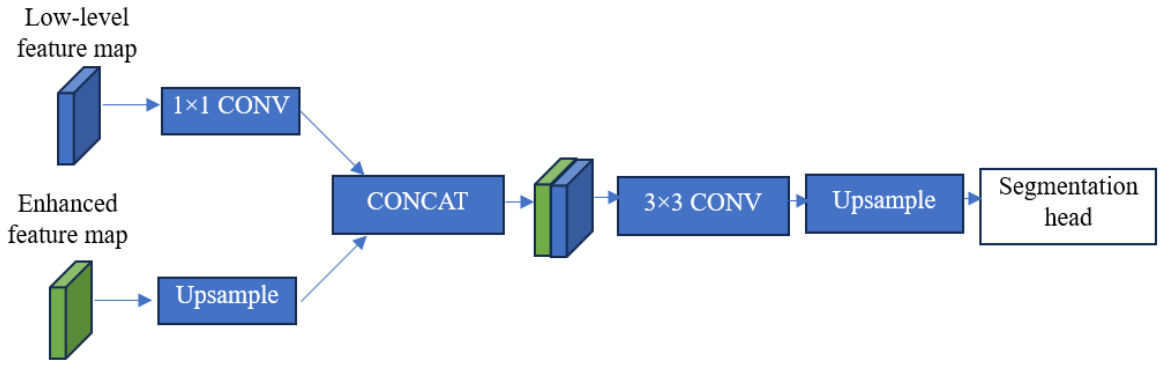

Fig. 3. The pipeline of the ASPP decoder.

C. Cropping

To extract high-resolution crack patches from the original image based on the coarse crack map, we first convert the coarse crack probabilities generated by the coarse network into a binary mask using a hard-thresholding method. This will produce a binary mask indicating the presence of cracks. To ensure that all potential details of the crack, even those that might have been slightly missed by the coarse network, are captured, we apply a dilation operation on the binary mask. This slightly enlarges the mask region. Next, we compute the bounding box for each contiguous region in the dilated binary mask. To generate these bounding boxes for crack regions, we select all the pixels with a corresponding density mask value of "1". We then merge the eight-neighbor connected pixels into a large candidate region. Finally, we use the circumscribed rectangle of the candidate region to crop the original image. This box encapsulates the region potentially containing the crack. Furthermore, we filter out crops with resolutions below the density threshold to eliminate noise and reject low-resolution patches. This step is crucial because crack segmentation doesn't perform well on low-resolution patches. After extracting the corresponding high-resolution patches from the original image, we feed each one into the fine network for detailed segmentation. As this network focuses only on smaller regions containing potential cracks, it can pay more attention to details, yielding better segmentation quality. Once the fine module processes the patches, it generates high-resolution segmentation for each patch. These segmented patches are then projected back to their original positions in the full-resolution image to obtain the final segmented result. By using this approach, the model benefit from both the global context provided by the coarse module and the local detail-centric approach of the fine module, ensuring accurate segmentation of cracks even in high-resolution images.

TABLE I. THE DETAILS OF RESNET-101 STRUCTURE USED AS THE SHARED ENCODER IN THIS PAPER

Layer type	Output size	Details
Input	H × W × 3	
Conv1	H/2 × W/2 × 64	7×7 convolution, stride 2
Max Pooling	H/4 × W/4 × 64	3×3 max pool, stride 2
Conv2_x (3 blocks)	H/4 × W/4 × 256	[1×1, 64], [3×3, 64], [1×1, 256] for each block
Conv3_x (4 blocks)	H/8 × W/8 × 512	[1×1, 128], [3×3, 128], [1×1, 512] for each block
Conv4_x (23 blocks)	H/16 × W/16 × 1024	[1×1, 256], [3×3, 256], [1×1, 1024] for each block
Conv5_x (3 blocks)	H/32 × W/32 × 2048	[1×1, 512], [3×3, 512], [1×1, 2048] for each block

D. Training Loss

In the domain of crack segmentation, the objective is to classify each pixel as either being part of a crack or not. This pixel-wise classification task can be effectively formulated and optimized using the Binary Cross Entropy (BCE) loss as follow:

$$L = \frac{1}{N}\sum_{ij} BCE(y_{ij}, \hat{y}_{ij}) \quad (1)$$

$$BCE(y, \hat{y}) = -ylog(\hat{y}) - (1-y)\log(1-\hat{y}) \quad (2)$$

where, y is the true label of the pixel, $\hat{y}$ is the predicted probability of the pixel belonging to the class labeled as foreground class.

The BCE loss quantifies the divergence between the predicted probability distribution and the actual distribution of the pixel labels. Specifically, for every pixel in the segmented image, the model predicts a probability score representing its confidence that the pixel belongs to the crack class. The BCE loss then computes the logarithmic difference between these predicted probabilities and the ground truth labels. When the model's prediction aligns closely with the actual label, the BCE loss approaches zero, indicating a perfect prediction. Conversely, if the model's prediction deviates significantly from the ground truth, the loss value increases. This property makes BCE loss particularly suitable for the crack segmentation task, as it penalizes misclassifications heavily, thereby driving the model to improve its pixel-wise classification accuracy. By minimizing the BCE loss during training, the segmentation model is guided to produce predictions that closely match the true crack structures in the images, leading to precise and reliable segmentation results. We use the BCE loss to both coarse and fine networks. The final loss L is the sum of the two:

$$L = \lambda_c L_c + \lambda_f L_f \quad (3)$$

where, λ_c and λ_f are coefficients for each loss, L_c is the BCE loss for the coarse network, L_f is the BCE loss for the fine network.

III. EXPERIMENTS

A. Dataset and Metrics

The proposed model is trained and evaluated on two public crack datasets: the CrackTree260 [23] and DeepCrack537 [24].

CrackTree260 is a dataset comprising 260 road pavement images, which is an extended version of the dataset from [23]. Images are captured using an area-array camera under visible-light illumination. This study utilized 200 of these images for training, 20 for validation, and 40 for testing. To enhance the training set, data augmentation techniques were employed. Specifically, each image was rotated at nine distinct angles ranging from 0 to 90 degrees in 10-degree increments. Following rotation, each image was then flipped both vertically and horizontally. From each flipped variant, five sub-images of 512×512 pixels were extracted – four from the corners and one from the center. As a result of this augmentation process, the training set accumulates a total of 35,100 images.

DeepCrack537 comprises 537 images, each having resolution of 544×384 pixels. These images depict a variety of cracks. Unlike other crack datasets, the diversity of cracks in DeepCrack537 is notable. Examples include top-down views, tilted views, cracks on both concrete and asphalt surfaces, variations in crack width from wide to thin, and instances where the cracks are partially occluded. For the purposes of this study, 300 images were utilized for training and 237 for testing.

For evaluation metrics, we use Precision (P), Recall (R), $F-measure$, mean intersection over union ($mIoU$) to evaluate the proposed model. Precision evaluates how many of the detected/segmented cracks (positive predictions) are actually real cracks.

$$P = \frac{TP}{TP+FP} \quad (4)$$

where, TP (True Positives) are the correctly detected cracks, FP (False Positives) are the wrongly detected cracks (i.e., detected cracks that are not real).

Recall measures how many of the actual cracks have been detected by the segmentation model.

$$R = \frac{TP}{TP+FN} \quad (5)$$

where, FN (False Negatives) are the actual cracks that the model failed to detect.

$F-measure$ is the harmonic mean of Precision and Recall. It provides a balance between the two. If either Precision or Recall is low, the $F-measure$ will also be low. It's especially useful when the class distribution is uneven.

$$F-measure = \frac{2\times P\times R}{P+R} \quad (6)$$

Mean intersection over union ($mIoU$) is a popular metric for segmentation tasks. It evaluates the overlap between the ground truth segmentation and the predicted segmentation.

$$mIoU = \frac{1}{2}\left(\frac{TP}{TP+FP+FN} + \frac{TN}{TN+FP+FN}\right) \quad (7)$$

B. Experimental Settings

The shared encoder based on ResNet-101 architecture is initialized with weights pre-trained on the ImageNet dataset [25] to take advantage of the extensive pretrained insights and rich feature representations it offers. The entire network was trained using the Adam optimizer [26] with a learning rate of 0.0001, which was reduced by a factor of 10 whenever the validation loss stabilized. Data augmentation techniques, including random rotations, zooms, and horizontal flips, were applied to prevent overfitting and enhance the model's generalization capabilities. The model is trained for 100k iterations with a batch size of 4. All experiments were conducted on a machine equipped with an NVIDIA RTX 4080 GPU, 64 GB RAM, and ran on the PyTorch framework.

C. Performance Evaluation

To demonstrate the effectiveness and superiority of the proposed method, we adopted eight existing and popular methods to compare to the proposed model, including Unet [6], TransUNet [27], FCNs [7], SegNet [5], and DeepCrack [28]. Table II shows results on the CrackTree260 dataset. From Table II, when examining the performance metrics of various models on the CrackTree260 dataset, the proposed model demonstrates superior performance across all metrics. With P of 0.892, R of 0.886, F-measure of 0.897, and mIoU of 0.894, the proposed model surpasses the other models in the ability to detect and segment cracks accurately. Notably, while DeepCrack comes closest to the proposed model with an F-measure of 0.852 and mIoU of 0.865, the proposed model still offers improvements, particularly in capturing the true positive rate as indicated by the highest recall. UNet and SegNet also demonstrate competitive results, with F-measures of 0.847 and 0.844 respectively. However, FCNs, with an F-measure of 0.463 and mIoU of 0.612, is the least effective among the mentioned models. TransUNet, despite being a transformer-based model, shows a relatively moderate performance with an F-measure of 0.771. Overall, the results suggest that the two-step approach of the proposed model is highly effective in detecting and segmenting cracks on the CrackTree260 dataset. For the DeepCrack537 dataset, the results are shown in Table III. From Table III, it's evident that the proposed model delivers the most impressive results in terms of crack detection and segmentation. With P of 0.891, R of 0.846, F-measure of 0.875, and mIoU of 0.878, the proposed model surpasses the other evaluated models. The DeepCrack model is the closest competitor with an F-measure of 0.847 and mIoU of 0.861, indicating a relatively high accuracy. SegNet also exhibits strong results with an F-measure of 0.840. In comparison, UNet and FCNs, while displaying reasonable performances, fall slightly behind with F-measures of 0.815 and 0.812, respectively. It is apparent from these results that the bi-level approach of the proposed model, involving a global context capture followed by focused high-resolution segmentation, is effective in the context of the DeepCrack537 dataset.

The visualization results of the model across datasets are depicted in Fig. 4. In this figure, the rows labeled "Input image" display images of various surfaces, each with distinct crack morphologies. The "Ground-truth labels" rows accurately depict the actual cracks present in the images. In contrast, the "Segmentation results" rows present the model's predictions. For most of the images, the segmentation results align closely with the ground-truth labels, indicating a high degree of accuracy in crack detection. This demonstrates the strength of our method in capturing both the global context through the coarse network and then refining the details through the fine network. Particularly in areas with intricate crack patterns or smaller fissures, the fine network provides superior results in segmenting these challenging features. However, there are instances, especially in the second set of images, where the segmentation results show a slightly broader crack outline than the ground-truth, suggesting a potential overestimation by the model in those cases. Overall, the visualization validates the effectiveness of our approach. While there are minor deviations in a few cases, the proposed pipeline demonstrates robust performance in segmenting cracks across diverse scenarios.

TABLE II. RESULTS ON THE CRACKTREE260 DATASET

Model	Metrics			
	P	R	F-measure	mIoU
UNet	0.860	0.834	0.847	0.861
TransUNet	0.797	0.746	0.771	0.803
FCNs	0.519	0.418	0.463	0.612
SegNet	0.851	0.837	0.844	0.858
DeepCrack	0.871	0.834	0.852	0.865
Proposed model	0.892	0.886	0.897	0.894

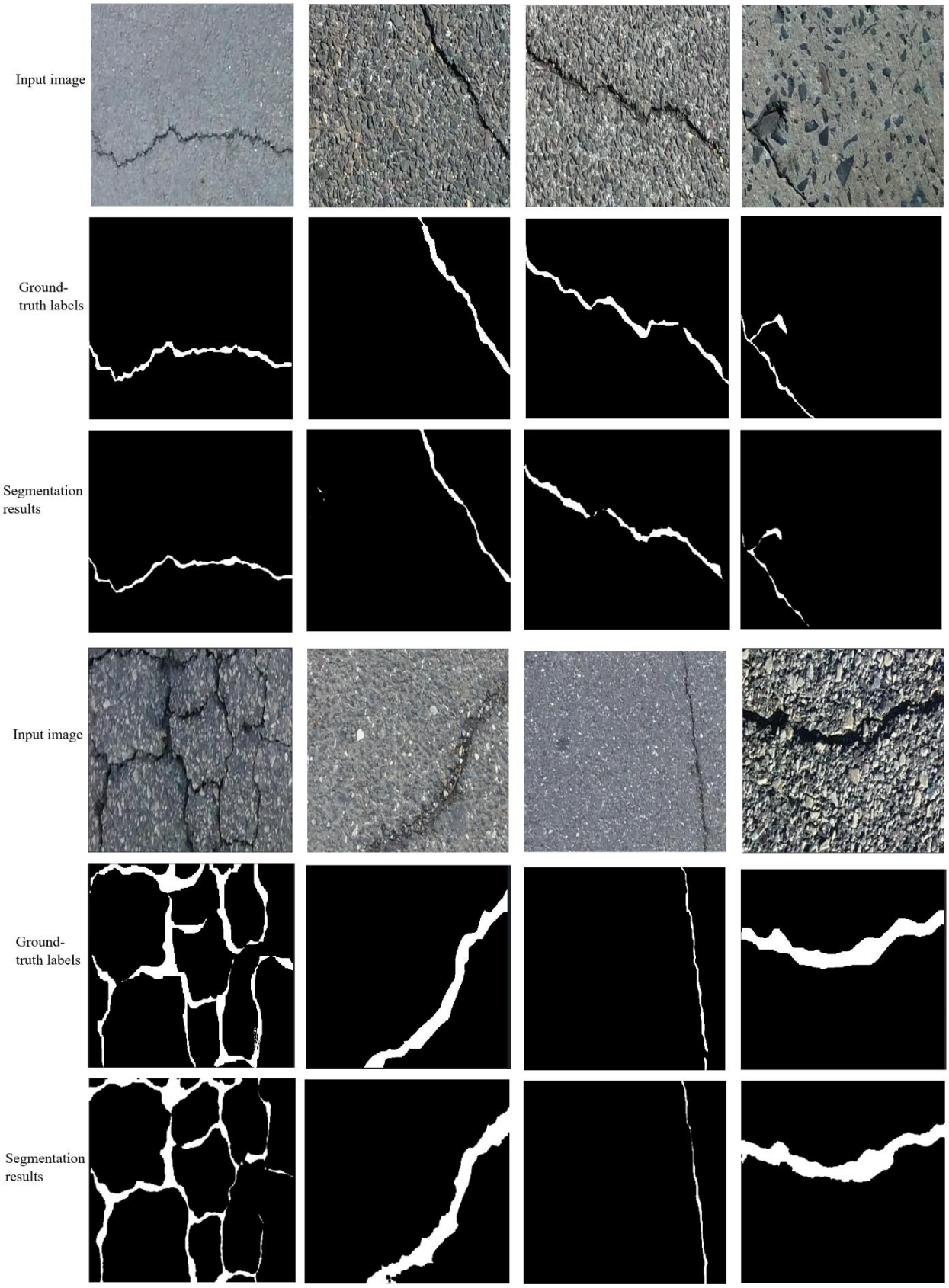

Fig. 4. Visualization results of the model on the datasets.

TABLE III. RESULTS ON THE DEEPCRACK537 DATASET

Model	Metrics			
	P	R	F-measure	mIoU
UNet	0.841	0.791	0.815	0.837
FCNs	0.829	0.796	0.812	0.833
SegNet	0.857	0.824	0.840	0.851
DeepCrack	0.876	0.819	0.847	0.861
Proposed model	0.891	0.846	0.875	0.878

IV. CONCLUSION

This study presents a dual-network model that optimally leverages the combined strengths of both a coarse and a fine network to enhance crack segmentation. Each network utilizes a shared encoder, with separate decoders tailored to their specific roles: the coarse network captures a holistic view of the image, emphasizing regions that potentially contain cracks, while the fine network focuses on these highlighted regions for precise high-resolution crack segmentation. The integration of the coarse network with the fine network was effective in addressing the challenges of crack detection. The experimental results on two public datasets underscore the robustness and effectiveness of the proposed approach in diverse scenarios. However, our model may struggle with extremely subtle cracks or those obscured by environmental factors, such as shadows or debris. Furthermore, the model's performance might vary depending on the quality and resolution of the input images. In future work, we plan to incorporate additional data augmentation techniques and explore the integration of advanced sensors for better crack detection in challenging conditions.

REFERENCES

[1] He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In *Proceedings of the IEEE international conference on computer vision*, pp. 2961-2969. 2017.

[2] Zhou, Bolei, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. "Semantic understanding of scenes through the ade20k dataset." *International Journal of Computer Vision* 127 (2019): 302-321.

[3] Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems* 28 (2015).

[4] Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. "Ssd: Single shot multibox detector." In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 21-37. Springer International Publishing, 2016.

[5] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." *IEEE transactions on pattern analysis and machine intelligence* 39, no. 12 (2017): 2481-2495.

[6] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234-241. Springer International Publishing, 2015.

[7] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431-3440. 2015.

[8] Chen, Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." *IEEE transactions on pattern analysis and machine intelligence* 40, no. 4 (2017): 834-848.

[9] Chen, Liang-Chieh, George Papandreou, Florian Schroff, and Hartwig Adam. "Rethinking atrous convolution for semantic image segmentation." *arXiv preprint arXiv:1706.05587* (2017).

[10] Ayomide, Kabirat Sulaiman, Teh Noranis Mohd Aris, and Maslina Zolkepli. "Improving Brain Tumor Segmentation in MRI Images through Enhanced Convolutional Neural Networks." *International Journal of Advanced Computer Science and Applications* 14, no. 4 (2023).

[11] Gunawan, Alexander Agung Santoso, Ilma Arifiany, and Edy Irwansyah. "Semantic segmentation of aerial imagery for road and building extraction with deep learning." *ICIC Express Letters* 14, no. 1 (2020): 43-51.

[12] Farooqui, Mehwash, Atta-ur Rahman, Roaa Alorefan, Mariam Alqusser, Lubna Alzaid, Sara Alnajim, Amal Althobaiti, and Mohammed Salid Ahmed. "Food Classification Using Deep Learning: Presenting a New Food Segmentation Dataset." *Mathematical Modelling of Engineering Problems* 10, no. 3 (2023).

[13] Mohan, Arun, and Sumathi Poobal. "Crack detection using image processing: A critical review and analysis." *alexandria engineering journal* 57, no. 2 (2018): 787-798.

[14] Adhikari, R. S., O. Moselhi, and A. Bagchi. "Image-based retrieval of concrete crack properties for bridge inspection." *Automation in construction* 39 (2014): 180-194.

[15] Prasanna, Prateek, Kristin J. Dana, Nenad Gucunski, Basily B. Basily, Hung M. La, Ronny Salim Lim, and Hooman Parvardeh. "Automated crack detection on concrete bridges." *IEEE Transactions on automation science and engineering* 13, no. 2 (2014): 591-599.

[16] Li, Haifeng, Dezhen Song, Yu Liu, and Binbin Li. "Automatic pavement crack detection by multi-scale image fusion." *IEEE Transactions on Intelligent Transportation Systems* 20, no. 6 (2018): 2025-2036.

[17] Yang, Fan, Lei Zhang, Sijia Yu, Danil Prokhorov, Xue Mei, and Haibin Ling. "Feature pyramid and hierarchical boosting network for pavement crack detection." *IEEE Transactions on Intelligent Transportation Systems* 21, no. 4 (2019): 1525-1535.

[18] Fei, Yue, Kelvin CP Wang, Allen Zhang, Cheng Chen, Joshua Q. Li, Yang Liu, Guangwei Yang, and Baoxian Li. "Pixel-level cracking detection on 3D asphalt pavement images through deep-learning-based CrackNet-V." *IEEE Transactions on Intelligent Transportation Systems* 21, no. 1 (2019): 273-284.

[19] Zhang, Zhengxin, Qingjie Liu, and Yunhong Wang. "Road extraction by deep residual u-net." *IEEE Geoscience and Remote Sensing Letters* 15, no. 5 (2018): 749-753.

[20] Liu, Huajun, Xiangyu Miao, Christoph Mertz, Chengzhong Xu, and Hui Kong. "Crackformer: Transformer network for fine-grained crack detection." In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3783-3792. 2021.

[21] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[22] Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801-818. 2018.

[23] Zou, Qin, Yu Cao, Qingquan Li, Qingzhou Mao, and Song Wang. "CrackTree: Automatic crack detection from pavement images." *Pattern Recognition Letters* 33, no. 3 (2012): 227-238.

[24] Liu, Yahui, Jian Yao, Xiaohu Lu, Renping Xie, and Li Li. "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation." *Neurocomputing* 338 (2019): 139-153.

[25] Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248-255. Ieee, 2009.

[26] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[27] Chen, Jieneng, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. "Transunet: Transformers make strong encoders for medical image segmentation." *arXiv preprint arXiv:2102.04306* (2021).

[28] Zou, Qin, Zheng Zhang, Qingquan Li, Xianbiao Qi, Qian Wang, and Song Wang. "Deepcrack: Learning hierarchical convolutional features for crack detection." *IEEE transactions on image processing* 28, no. 3 (2018): 1498-1512.