IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 January 2024, accepted 3 February 2024, date of publication 22 February 2024, date of current version 8 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3368376

RESEARCH ARTICLE

On Enhancing Crack Semantic Segmentation Using StyleGAN and Brownian Bridge Diffusion

COLLINS C. RAKOWSKI AND THIRIMACHOS BOURLAI
College of Engineering, University of Georgia, Athens, GA 30602, USA
Corresponding author: Thirimachos Bourlai (Thirimachos.Bourlai@uga.edu)

ABSTRACT Inspection for cracks is an essential yet labor-intensive aspect of maintenance for structures in active service bridges. Deep learning networks, combined with an abundance of segmented image data representing various types of cracks, enable the development of a computer vision-based solution. Often, segmentation data is scarce and requires a great deal of time to annotate. This paper introduces a novel approach to structural crack detection using synthetic data generation and advanced semantic segmentation models. We employ StyleGAN3 and the Brownian Bridge Diffusion Model (BBDM) to create a diverse and realistic dataset of synthetic structural crack images, addressing the critical challenge of creating segmentation data in training deep learning models for crack detection. Our methodology is based upon the DeepLabV3+, i.e., a semantic segmentation architecture that builds on DeepLabv3 by adding a simple yet effective decoder module to enhance segmentation results. The original DeepLabV3+ model is insufficient and thus, we first perform a meticulous hyperparameter tuning, which is responsible for about a 10% increase in overall performance. Next, we generate multiple image-to-image translations with BBDMs synthetic datasets and pair them with a set of fine-selected data augmentation techniques, including motion, zoom, and defocus blur, to improve crack segmentation performance. When compared to the state-of-the-art latest work on the same database that achieved an accuracy of 61.49%, our proposed work attains a Mean Intersection over Union (MeanIoU) accuracy of 65.62% through ensemble modeling on multiple synthesized datasets, employing a majority voting strategy. We also showcase the potential of diffusion models in synthetically generated datasets that elevate semantic segmentation accuracy and introduce blur augmentation as a viable technique for enhancing model robustness. The results indicate that our approach not only surpasses conventional methods in terms of MeanIoU but also offers a new avenue of research into diffusion-model-based synthetic image generation for improved semantic segmentation performance.

INDEX TERMS StyleGAN, DeepLabV3+, synthetic data generation, hyperparameter tuning, Brownian bridge diffusion model, semantic segmentation, data augmentation, ensemble modeling, structural crack detection.

I. INTRODUCTION

Structural safety is of paramount importance in the industrial community, particularly in the context of in-service bridges. Ensuring the integrity of these structures is vital to preventing accidents and maintaining the safety of people and goods that rely on them. Among the many challenges faced in this domain, the detection of structural cracks stands out as a critical concern. The early identification and monitoring of cracks can significantly contribute to the longevity and reliability of these infrastructure assets. Conventional methods use visual inspection (traditional visual inspection methods involve human inspectors visually examining bridge components for cracks and anomalies), Non-Destructive Testing (NDT), i.e.,

The associate editor coordinating the review of this manuscript and approving it for publication was Xiong Luo.

techniques such as ultrasonic testing, radiography, and magnetic particle inspection, acoustic emission testing, vibration analysis, and other [1]. All the aforementioned approaches are limited due to various reasons. Visual inspections, being subjective and reliant on individual expertise, may result in inconsistencies. Non-destructive testing methods often face limitations in coverage and accessibility, proving challenging for comprehensive inspections. Moreover, these traditional approaches can be time-consuming, labor-intensive, sensitive to environmental conditions, and may lack the sensitivity to detect early-stage or hidden cracks. In contrast, image processing and deep learning offer advantages in terms of automation, objectivity, continuous monitoring, and adaptability, making them valuable alternatives or complementary solutions to traditional crack detection techniques in bridge maintenance. Even though such a solution is attractive, this task is far from straightforward, as it involves both data collection and annotation.

The original approach to structural crack detection using deep learning often requires extensive data collection, which can be time-consuming and resource-intensive. Gathering a substantial number of crack images is essential for training deep learning-based crack detection systems; this approach has emerged as a promising solution to the problem. These novel deep learning architectures have the potential to autonomously learn to detect cracks, but they demand large datasets to acquire a comprehensive understanding of the diverse features and patterns associated with real-world cracks [2]. Moreover, the complexity of the task is compounded when we introduce variable noise factors from data collection such as motion blur, defocus blur, and zoom blur.

While significant strides have been made in the field of deep learning-based crack detection, one primary research paper has pioneered a novel approach [3]. This paper presents a diffusion model-based method for creating a synthesized crack image dataset with pixel-wise annotations. This innovative approach complements traditional data augmentation techniques and offers a different perspective on addressing the challenges of creating segmentation data.

Jin et al. [3] employed a Deep Convolutional GAN (DCGAN) to generate synthesized crack annotations while using a Pixel2Pixel [4] model to create corresponding synthesized crack images. They meticulously documented the learning process of these GANs during training, revealing how they adapt to generate synthesized images that mimic real-world crack patterns. Importantly, they also conducted a comparative study to assess the performance of Deep Neural Networks (DNNs) trained with this synthesized dataset. Their research explores different approaches to using synthesized and real crack images for training DNNs, with a notable finding that pre-training with synthesized images followed by fine-tuning with real images outperforms direct mixing of both types.

As we discuss in the literature review section, there is extensive research showing the capabilities of DNNs in crack segmentation, but a gap remains for us to explore the capabilities of diffusion-based synthesized datasets.

Building upon the foundation laid by the aforementioned research, our work endeavors to take diffusion-based synthetic crack data generation and semantic segmentation to new heights through newer hyper-tuned models, augmentation techniques, and ensemble modeling. While the most recent paper [3] made significant strides in introducing GANs for generating synthetic crack data and their application in segmentation, there are a set of limitations that we address. These limitations include limited hyperparameter tuning, the absence of performance ensemble modeling, insufficient augmentation studies, and a lack of exploration into diffusion-based synthetic image data generation approaches.

A. CONTRIBUTIONS

Our approach introduced in this paper makes several significant contributions to the field of structural crack detection using deep learning and synthetic data generation:

1) *Novel Approach to Synthetic Data Generation*: A major contribution of this research is the implementation of StyleGAN3 and the BBDM for generating synthetic crack images. Harnessing StyleGAN3, we generate mask annotations that provide precise information about the crack regions. Simultaneously, we employ BBDM to create synthetic real images corresponding to these mask annotations. This novel approach successfully addresses the challenge of creating segmentation data, which is a critical limitation in training deep learning models for accurate and reliable crack detection.
2) *Enhanced Performance through Hyperparameter Tuning of DeepLabV3+*: The research demonstrates the substantial impact of hyperparameter tuning on the performance of the DeepLabV3+ model. We experiment with various hyperparameters, like changing backbones and optimizers, and incorporate data augmentation with the goal of enhancing semantic segmentation accuracy. Meticulous optimization of various hyperparameters has led to a 9% improvement in model accuracy, underscoring the importance of fine-tuning in semantic segmentation tasks.
3) *Blur Data Augmentation Technique in the Context of Synthesized Dataset*: The study introduces blur data augmentation techniques, including motion, zoom, and defocus blur, specifically applied to synthesized datasets. This method enhances the robustness of the models, making them more adept at handling real-world variations in image quality and contributing to the improvement in segmentation performance.
4) *Ensemble Modeling on Augmented Synthesized Datasets*: A pivotal aspect of the research is the application of ensemble modeling techniques to models trained on data-augmented synthesized datasets. By leveraging the strengths of multiple models through ensemble modeling, particularly with majority voting, the study achieves notable enhancements in

MeanIoU accuracy. This demonstrates the efficacy of ensemble modeling in extracting maximum value from synthesized, augmented data.

5) *Diffusion-Based Synthetic Image Data Generation*: The application of BBDM in the context of synthetic data generation represents a pioneering exploration in the field. This research showcases the capabilities of diffusion models in generating high-fidelity, diverse crack images, paving the way for broader applications of diffusion-model-based image generation in machine learning and semantic segmentation.

We present a comprehensive solution that tackles the challenge of generating segmentation data in structural crack detection through the innovative use of ensemble learning. By integrating multiple models, each trained on distinct datasets comprising both real and synthetically generated images, we significantly enhance the performance of crack detection systems using ensemble modeling. This ensemble approach not only diversifies the training data but also combines the unique strengths of individual models to improve overall accuracy and robustness.

Our results demonstrate the highest recorded performance on the Bridge Crack Library (BCL) dataset [5], achieving a MeanIoU of 65.62%, which sets a new benchmark in structural crack detection. This breakthrough illustrates the efficacy of our method in addressing the perennial issue of limited data availability, a hurdle that has long impeded progress in applying deep learning techniques to real-world structural safety challenges. By leveraging the synergy of multiple models trained on a rich mix of real and augmented synthetic data, we provide a path forward for the effective application of deep learning in scenarios where high-quality, segmented data is scarce or difficult to obtain. Our approach not only advances the field of structural crack detection but also offers a blueprint for tackling similar challenges across various domains where data limitations persist.

II. LITERATURE REVIEW

A. NETWORK TUNING AND ARCHITECTURE

The recent advancements in network tuning and architecture for semantic segmentation play a pivotal role in enhancing the accuracy and efficiency of structural crack detection. This section delves into various innovative approaches and modifications to existing models, focusing on their application in the precise identification and segmentation of structural cracks. Several subsequent papers have contributed to the advancement of semantic segmentation models crucial for structural crack detection.

DeepLabV3+ [6], with its integration of spatial pyramid pooling and depthwise separable convolution in both ASPP and decoder modules, has demonstrated notable performance on benchmark datasets, making it a preferred choice for crack segmentation. Further enhancements, such as those by Fu et al. [7], who introduced a densely connected ASPP to DeepLabV3+, and comprehensive evaluations by Wang et al. [8] across various crack detection domains, underline the model's adaptability and efficacy.

The efficacy of different backbone architectures in DeepLabV3+ has been explored, with studies such as Atik et al. [9] assessing the performance across ResNet [10], Xception [11], and MobileNetV2 [12] backbones, and Nie et al. [13] comparing ResNet, DenseNet, and EfficientNet [14] for defect detection. These studies highlight the impact of backbone choice on segmentation performance. Beyond traditional architectures, Ye et al. [15] introduced the Pruned Crack Recognition Network (PCR-Net), optimized for mobile devices, showcasing the potential for on-site, efficient crack detection utilizing edge computing and deep learning.

Ensemble learning has emerged as a successful strategy for improving semantic segmentation outcomes. Fan et al. [16] and Hirata and Takahashi [17] demonstrated the effectiveness of combining specialized CNNs for fine crack structure detection, achieving high precision and recall. Similarly, Kailkhura et al. [18], and Rodriguez-Lozano et al. [19] utilized ensemble methods, such as averaging and majority voting with pre-trained CNN models, to significantly uplift model accuracy and performance metrics in concrete crack and road pavement detection. Li et al. [20] study addresses the challenge of tunnel crack detection, an area that's crucial for maintaining tunnel safety but presents unique challenges compared to more studied pavement crack detection. Additionally, we saw Maarouf and Hachouf [21] proposed a transfer learning-based ensemble deep learning approach to improve segmentation using four pre-trained deep CNN architectures AlexNet [22], GoogleNet [23], VGG16 [24], and ResNet50 for increased segmentation performance. Drawing inspiration from the aforementioned studies, this research has employed a strategic approach to ensemble learning to enhance crack detection accuracy.

These contributions collectively underscore the dynamic nature of network architecture and tuning in semantic segmentation, emphasizing the role of DeepLab enhancements, backbone diversity, and ensemble methods in pushing the boundaries of crack detection technology.

B. MODELS FOR CRACK SEGMENTATION

The field of crack segmentation has witnessed substantial advancements through the application of deep learning, particularly Convolutional Neural Networks (CNNs), for structural integrity assessment. Various models have been developed to address the complexities of crack detection across different surfaces and contexts.

Central to these advancements are U-Net and its enhancements, which have become foundational in the development of crack segmentation models. The U-Net architecture, introduced by Ronneberger et al. [25], has been pivotal in this evolution, with further contributions by Tang et al. [26] with Crack Unet, Han et al. [27] with CrackW-Net, and

Liu et al. [28] with YoloV3 [29], each demonstrating significant improvements in segmentation accuracy and operational efficiency in diverse crack domains.

Beyond U-Net-based models, the field has seen the introduction of several innovative architectures. For instance, Lau et al. [30] propose a network incorporating a pre-trained ResNet-34, and Wang and Su [31] utilize a Pyramid Attention Network which employs Densenet 121 [32]. Additionally, models like CrackU-net by Huyan et al. [33], SDDNet by Choi and Cha [34], and RUC-Net by Yu et al. [35] further expand the toolkit available for crack segmentation, each bringing unique enhancements to tackle the intricacies of crack detection.

These models have proved to be very successful in pixel-wise pavement segmentation [35], [36], [37], [38], [39], [40], [41], [42]. In another similar domain, Zhang et al. [43] and Shi et al. [44] assessed their models by detecting cracks in roads.

The contributions of DeepCrack by Liu et al. [45], offering a hierarchical feature learning architecture, and CrackSegAN by Pan et al. [46], employing a GAN-based approach, underscore the importance of hierarchical feature learning and generative models in achieving high precision in crack segmentation. Furthermore, Ali et al. [2] emphasize the significance of evaluation metrics such as accuracy, precision, recall, and F1-Score, particularly in the context of imbalanced datasets, highlighting the necessity of balanced detection sensitivity and specificity.

In a primarily CNN-dominant field, diffusion-based methods are establishing themselves as candidates for enhancing image segmentation tasks. Yu et al. [47] explored diffusion-based data augmentation for nuclei image segmentation. The study is pioneering in applying a diffusion-based method for data augmentation in the context of nuclei segmentation. This method is used to generate synthetic, labeled images to enhance training datasets. By synthesizing additional training data, this approach effectively circumvents the limitations imposed by the need for extensive manual labeling. The research demonstrates that augmenting just 10% of a labeled real dataset with these synthetic samples can yield segmentation results comparable to a fully supervised baseline. Dhariwal and Nichol [48] demonstrated the superiority of diffusion models when compared to deep learning models. Dhariwal and Nichol's paper is a landmark in the field of artificial intelligence and machine learning, specifically in the sub-domain of generative modeling. Together, these papers demonstrate the potential of diffusion models in machine learning. Diffusion-based methods have shown promise in enhancing image segmentation tasks, which inspired their incorporation into this study to explore new frontiers in image analysis and data synthesis. As diffusion models represent a cutting-edge advancement in the field of image generation, our study is among the pioneering efforts to leverage these models for creating diffusion model-based image datasets specifically tailored for semantic segmentation tasks.

C. DATA SYNTHESIS IN VARIOUS DOMAINS

The ability to generate synthetic images not only addresses the growing demand for high-quality, diverse datasets but also presents a solution to the challenges of data scarcity and privacy concerns. From healthcare to environmental studies, the creation and manipulation of image data through advanced computational methods have opened new avenues for research and application. Data synthesis has shown promise in various domains.

Ding et al. [49] and Fetty et al. [50] have made notable strides in medical imaging by introducing synthetic datasets for pathological and radiological images, respectively, utilizing StyleGAN variants and a Nuclei Annotator (NA) using HoVer-Net [51] for their generation and segmetation. This approach mitigates the reliance on extensive, labor-intensive human annotations and addresses data availability issues due to privacy concerns, showcasing the potential of generative adversarial networks (GANs) in medical data synthesis.

Further advancements by Karras et al. [52] with StyleGAN3 aims at enhancing image quality by eliminating aliasing, indicating a leap in the realism of generated images. Similarly, Xu et al. [53] extend the application of synthetic data generators to scenarios such as dam surface crack detection. Bartz et al. [54] trained a StyleGAN model to synthesize historical document images. This synthetic dataset was used to train various semantic segmentation networks, such as Doc-UFCN [55], EMANet [56], and TransUNet [57], tailored for specific tasks like line segmentation in historical documents. These studies highlight the versatility of synthetic data in training machine learning models across varied applications.

Innovative approaches by Rombach et al. [58] and Li et al. [59] introduce Latent Diffusion Models (LDMs) and the Brownian Bridge Diffusion Model (BBDM) for efficient high-resolution image synthesis and image-to-image translation. These models represent significant advancements in reducing computational demands and establishing direct mappings between image domains without conditional generation processes, showcasing state-of-the-art performance in a range of synthesis tasks.

Collectively, these studies highlight the effectiveness of advanced generative models like StyleGAN and diffusion methods in creating detailed, accurate synthetic datasets. The success of these approaches in diverse applications underscores the potential of data synthesis to overcome the limitations of traditional data collection and annotation methods, offering new avenues for research and application in the realm of image generation and manipulation.

D. SYNTHESIS OF STRUCTURAL CRACK DATASETS FOR ENHANCED DETECTION

Structural safety in the industrial sector, especially for in-service bridges, is paramount, necessitating efficient crack detection methods. However, conventional approaches

demand extensive data collection, which often results in data scarcity.

Zhai et al. [60] generated 3D synthetic data to improve crack identification performance, and Pei et al. [61] proposed a method for the virtual generation of pavement crack images based on an improved deep convolutional generative adversarial network (DCGAN).

Ye et al. [62] presented benchmark datasets for evaluating crack detection algorithms with the creation of the Bridge Crack Library Dataset. This dataset is the foundation upon which our data is generated. Shortly after, pioneering work by Jin et al. [3] introduced generative adversarial networks (GANs) for synthesizing crack datasets with pixel-wise annotations, addressing the issue of data scarcity. They propose a method that closely mirrors ours, where they establish a pipeline of GANs for data synthesis. They used Pix2Pix for synthesized mask image annotations and DCGAN for synthesized crack image generation which led to the development of Bridge Crack Library 2.0 (BCL 2.0) [63], a fully synthesized dataset derived from the original BCL. This parallel in methodology ensures that the comparison of segmentation performances between BCL 2.0 and our generated datasets is based on a level playing field, with both datasets benefiting from synthetic generation. However, their segmentation performance was evaluated on different architectures than our own. They used PCR-Net, DeepLab, U-Net, and FCN [64] to assess their BCL 2.0 dataset. Their best model, PCR-Net, achieved a remarkable MeanIoU of 74.34%, surpassing models trained on real images. Moreover, they found that pre-training with synthesized images, followed by fine-tuning with real images, proved more effective than mixing both the real and synthesized datasets. It is important to note that the dataset used by Jin et al. for their testing, comprising their own set of 1000 images, is not publicly available, underscoring a limitation in the generalization of their findings. This discrepancy in dataset size could have implications for the training depth and model robustness because, in our research, 1000 images of the training set had to be set aside for testing.

In our study, we expand upon this approach by introducing newer GANs, StyleGAN3, and a diffusion model, BBDM, for image-to-image translation. Similarly, we generate a synthetic dataset but take it one step further by applying blurring. To assess our dataset, DeepLabV3+ was chosen because of the accessibility and proven performance of this architecture in semantic segmentation tasks. The PCR-Net architecture, while yielding promising results, was not readily available for implementation. Therefore, we opted for the next most viable and accessible model, DeepLabV3+, which is well-regarded in the field. This choice is validated by the favorable results we achieved, particularly in comparison with Jin's [3] findings. Our model not only demonstrated competitive performance but also highlighted the capability of widely accessible models like DeepLabV3+ to achieve high-precision results in complex segmentation tasks.

III. METHODOLOGY

We present a novel methodology for synthetic data generation that circumvents the traditional data annotation process, which can be very timely. Our methodology comprises a multi-step process that can be found in Figure 3.

In the first step, we employ StyleGAN3 to train on crack annotation images sourced from the BCL mask annotations. The BCL mask annotations were created by Ye et al. [62]. Refer to Figure 2 and to the Bridge Crack Library Dataset section for more details. This initial GAN-based step generates annotation mask images, which serve as the foundation for our subsequent synthetic data generation. StyleGAN3 is particularly well suited for the precise task of generating mask images because of its ability to reproduce fine details, which makes it ideal for reproducing intricate and varied patterns of cracks. In addition, StyleGAN3 is capable of producing a diverse array of cracks, which is crucial for training robust semantic models. We introduce a BBDM into our pipeline, enabling the translation between domains, specifically converting annotation mask images into realistic synthetic structural crack images. This translation step is critical for generating synthetic data that closely resembles actual crack images in both texture and structure. Unlike traditional methods that rely on conditional generation processes, BBDM directly translates between two image domains through a stochastic process. BBDM's ability to generate high-fidelity images ensures that the synthesized crack images are realistic. This is the first time BBDM has been used in the context of crack image data generation and is also a first for diffusion-based crack image data generation. In step two, we hyper-tune the DeepLabV3+ model using the data in Table 2 and then use the newly hyper-tuned DeepLabV3+ model to train on the blur-augmented synthesized datasets. Finally, we ensemble the trained model predictions together using majority voting to garner the best MeanIoU performance.

A. CREATING SYNTHESIZED DATASETS

1) STYLEGAN3

StyleGAN3 is an advanced variant of GANs known for its ability to generate high-quality and diverse images. StyleGAN3's role in our methodology is to learn and replicate the intricate patterns and characteristics present in crack annotation images sourced from the BCL dataset. These annotation mask images, which typically depict the outlines and boundaries of structural cracks, are critical for subsequent steps in our pipeline.

The choice of StyleGAN3 for generating synthetic mask annotations of cracks is rooted in its advanced architectural features. Unlike its predecessors, StyleGAN3 introduces several significant innovations that enhance the quality and stability of the generated images. A notable aspect of StyleGAN3 is its redesigned generator architecture, which incorporates an adaptive discriminator augmentation mechanism. This mechanism is particularly effective in

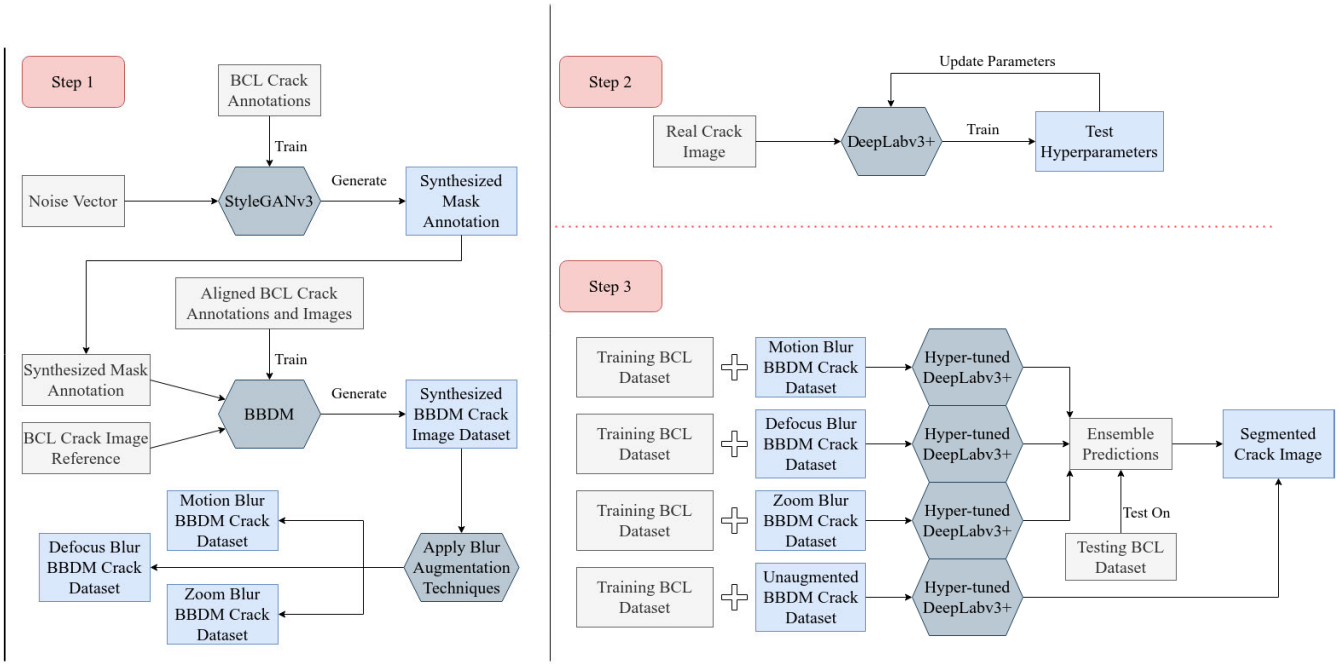


FIGURE 1. Overall methodology schematic: A 3-step process was implemented to produce a synthesized dataset with BBDM in the pixel and latent spaces. In step one, StyleGAN3 is trained on the ground truth mask annotation images from the BCL dataset. Once trained, a noise vector is passed into the model, and a new randomly generated mask annotation image is formed. The pixel and latent space BBDM were trained on the entire BCL dataset. The newly generated mask annotations were randomly paired with a BCL crack image for inferencing purposes. The result is a synthesized BBDM crack image dataset. Three additional blur-augmented datasets were created by copying the original synthesized BBDM crack dataset. In step two, DeepLabV3+'s hyperparameters are tuned to improve overall segmentation accuracy. In step three, 8 different datasets, 4 from the pixel space BBDM and 4 from the latent space BBDM, were trained using the new hyper-tuned DeepLabV3+. All the predictions from the models trained on the datasets with blur augmentation were ensembled together to generate a final segmented crack image.

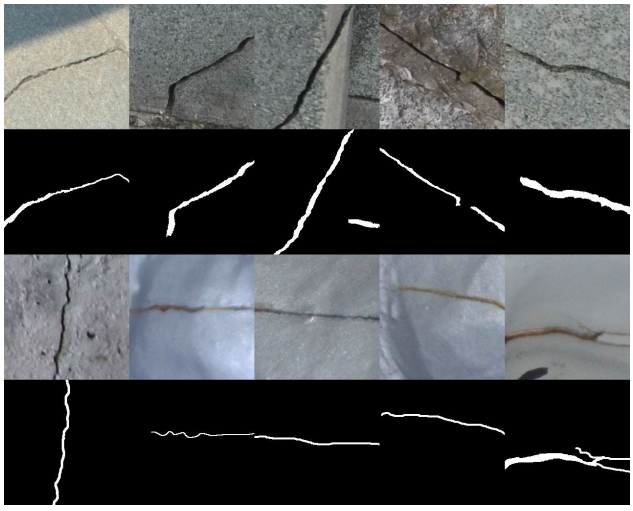


FIGURE 2. Bridge crack library sample images. The figure shows a selection of images from the Bridge Crack Library, where each picture captures different types of cracks found in materials like stone, concrete, and metal. These pictures, all with the same size of 256 by 256 pixels, are part of a big collection gathered from inspecting over 50 bridges. This collection includes 5,769 non-steel crack images (as seen in row one), 2,036 steel crack images (as seen in row three), and 3,195 images identified as noise. Each of these images also has an annotated pair as seen in row two and four.

preventing overfitting, a common challenge when training on a relatively small, specialized dataset like the BCL. The network also includes refinements in the mapping and synthesis networks, which contribute to the generation of more detailed and varied images, enhancing the model's ability to produce diverse and realistic crack patterns.

2) TRAINING STYLEGAN3

The training process for StyleGAN3 is carried out over multiple days using the following hardware: 4 NVIDIA A6000 GPUs on Ubuntu. All images are processed at a resolution of 256 by 256 pixels, and no preprocessing was necessary to train on the BCL dataset. The batch size is set to 16 to optimize GPU usage. The Gamma value is set to 8 to regularize the trade-off between fidelity and variety of the synthesized images. Specifically, the gamma parameter in StyleGAN3 plays a pivotal role in balancing the fidelity and diversity of generated images. We choose a gamma value of 8. For approximately 10,000 epochs, we observe a progressive refinement in the quality of the generated images. Initially, the images bore only a rudimentary resemblance to crack patterns, but the final images exhibit a more defined and realistic appearance, closely mimicking various types of crack formations. In Figure 3, the sample synthesized mask image results show a wide variety of cracks.

All the cracks are textured and resemble many different types of damaged surfaces, including thin, thick, bifurcated, rugged, and smooth cracks. All the cracks manifested in the same orientation. To diversify the generated images,

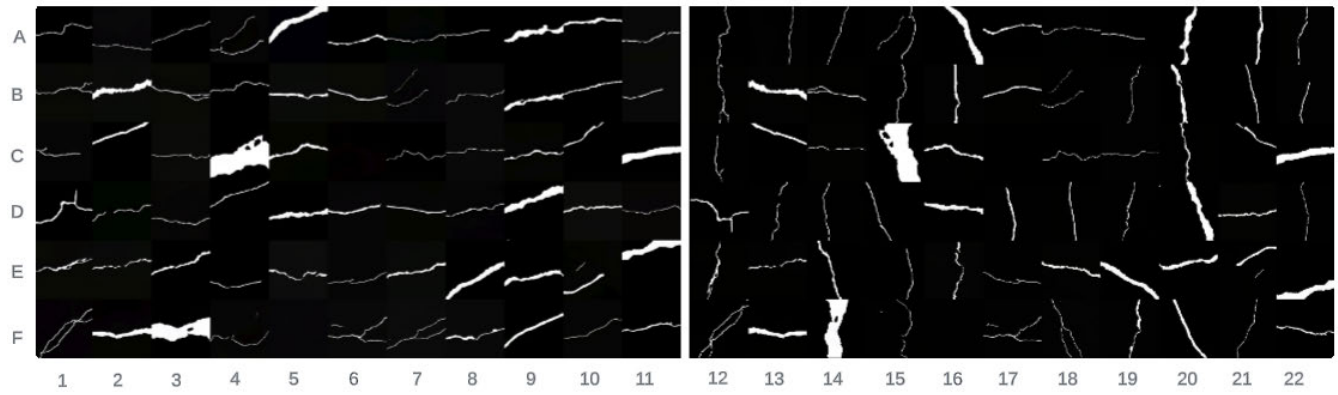


FIGURE 3. StyleGAN3 generated and rotated results. Each tile displays a distinct crack formation with varying orientations and morphologies, ranging from linear to branched, from fine to wide cracks, and from smooth to rugged. The diversity in crack presentations simulates a wide array of real-world conditions. The rotation of crack images in columns 12 through 22 exemplifies the model's ability to depict varied crack orientations, enriching the dataset's diversity. Generating cracks uniformly across a single axis simplifies dataset compilation and ensures the model's impartiality to crack orientation, leading to a more reliable crack segmentation model.

we implemented a pragmatic approach involving data augmentation. We applied random flips and rotations (0, 90, 180, and 270 degrees) to the input images. This strategy introduced the necessary variability and randomness, effectively ensuring a diverse representation of crack types, as seen in Figure 3. This augmentation technique was applied to all the synthesized mask annotations. The resulting synthetic crack images exhibited a remarkable range of textures and forms, closely resembling real-world crack patterns, which demonstrates StyleGAN3's capability to produce a comprehensive set of synthetic annotations that can robustly support various semantic segmentation tasks.

3) BBDM

The BBDM presents a novel architecture for image-to-image translation, which is particularly relevant for our project focused on generating synthetic crack images. This model stands apart from conventional diffusion models because of its unique handling of the image translation process. BBDM has never been used in the context of crack image synthesis.

At the core of BBDM, the stochastic Brownian Bridge process is defined by its conditional distribution that is dependent on both the starting and ending states of the diffusion journey. This dual anchoring is pivotal in guiding the model to adhere closely to the target domain while maintaining a structured and relevant approach to image translation. Unlike conventional models that end in random noise, the Brownian Bridge's endpoint is a known target state, which is instrumental in ensuring the generated images align precisely with the desired outcomes, such as accurate representations of crack patterns. The target state in this case will be the BCL dataset. This bidirectional methodology ensures that the translation is not only visually appealing but also accurately represents the target domain. In the forward process, BBDM offers the marginal distribution at each step, enabling a granular understanding of the image transformation. Conversely, the reverse process in BBDM starts with a conditional input, strategically aligning the model's output with the ultimate goal of realistic image synthesis. In the context of translating crack annotations to realistic crack images, this direct learning approach is vital for maintaining precision and detail, crucial elements in applications like crack detection.

When considering the application of BBDM in generating synthesized datasets, the distinction between pixel space and latent space BBDM becomes significant. While the latent space BBDM works in a Vector Quantize Generative Adversarial Networks (VQGAN) compressed feature space emphasizing efficiency and the capability to generate a variety of textures, the pixel space BBDM operates directly on image pixels, offering high precision in crack generation.

4) TRAINING BBDM

In our research, we explored the capabilities of both the pixel space and the latent space of BBDM for generating a synthesized dataset. We train both models on the entirety of the images and annotations from the BCL dataset. Then, we aligned the newly synthesized StyleGAN3 mask images with a random BCL real image for domain inferencing to generate the fully synthetic dataset.

The latent space Brownian Bridge Diffusion Model (LBBDM) was trained with a focus on efficiency and generalization. This model underwent a smaller training regime, spanning 400,000 steps. The batch size is kept consistent at 8, and like the pixel space model, the optimizer used is Adam with a similar learning rate and beta1 settings. LBBDM also uses VQGAN for its ability to compress high-dimensional data into a latent space while preserving crucial image features, and it implements a UNet architecture tailored to the 64 × 64 resolution for faster training and inferencing. VQGAN was pre-trained on VQ-f4.[1]

[1]VQ-F4 can be retrieved from this official repository https://github.com/CompVis/latent-diffusion

The pixel space BBDM is configured to effectively learn and translate crack patterns at the pixel level. It was trained for just over 600,000 steps on one GPU. Key training parameters include a batch size of 8. The model's optimizer is set to Adam, with a learning rate of 1.e-4 and a beta1 parameter of 0.9. The learning rate scheduler has a decay factor of 0.5 and patience of 3,000 steps, facilitating a dynamic adjustment of the learning rate to optimize the training process. The model also employs an Exponential Moving Average (EMA) with a decay of 0.995, starting at step 30,000. The architecture of the model, characterized by its UNet parameters, is designed to handle the 256×256 resolution images with specific attention to details in the crack patterns.

5) GENERATING MULTIPLE BLUR-AUGMENTED SYNTHETIC DATASETS

Following the training phase of the BBDM, we generated several synthetic datasets by using 200 sampling steps during the inference stage. The real crack images from the BCL dataset and the synthesized annotation images produced by StyleGAN3 are used as the domain reference images. Each dataset contains 7800 synthesized images, along with a set of real images from the BCL dataset. We produced four different datasets for both the pixel space and latent space models. To add variety, three out of the four datasets were modified with different types of blur: motion blur, zoom blur, and defocus blur. The goal of these augmented datasets is to attack some of the problems with the BCL dataset. The fourth dataset was kept as originally synthesized for baseline comparison.

There are distinct characteristics between the pixel space and latent space images generated by the BBDM. The pixel space approach demonstrated high fidelity in reproducing crack annotations accurately; however, it often struggled with generating realistic textured backgrounds, as seen in Figure 4a. This resulted in backgrounds that occasionally appeared washed out, even though the cracks themselves were well-rendered. The precision of crack representation in pixel space images is noteworthy, as it aligns closely with the detailed annotations but is sometimes limited by less convincing textural quality.

In contrast, the LBBDM showcased its strength in generating more diverse and realistic backgrounds, as demonstrated in Figure 4e. This enhancement in background texture contributes significantly to the overall realism of the images. However, this advantage comes with a trade-off. The latent space model tended to generate cracks that did not adhere as closely to the annotated crack patterns. The discrepancies in crack representation in the latent space images suggest a divergence from the source annotations, highlighting the model's focus on broader image context rather than precise detail replication.

Incorporating different types of blurs as data augmentation techniques plays a crucial role in enhancing the adaptability and robustness of our model, particularly in the context of the BCL dataset. Imgaug, a publicly available data augmentation tool created by Michaelis et al. [65], was used to apply different magnitudes and types of blurring. Refer to Figure 4.

Motion blur, which replicates the effect of rapid movement, is instrumental in preparing the model for scenarios involving moving subjects or capturing devices. A random severity ranging from 3 to 12 was applied to the pixel and latent BBDM datasets. This type of augmentation is particularly beneficial for analyzing images of cracks that are captured in motion, such as those taken from moving vehicles or through dynamic monitoring systems, ensuring that the model maintains consistent crack detection performance even in motion.

Similarly, zoom blur is employed to mimic the effect of rapid changes in focal length, challenging the model to maintain its accuracy despite variations in image focus. This augmentation is significant for images captured from varying distances or during swift zooming actions, a common occurrence in field surveys. A magnitude of one is applied to the zoom blur BBDM synthetic datasets.

Defocus blur, on the other hand, simulates scenarios where the image is not perfectly focused. Again, a severity of one is applied to the defocus blur BBDM synthetic datasets. This type of blur presents a challenge for the model to accurately recognize and segment cracks, even when the overall sharpness of the image is compromised.

For each synthesized dataset, the set was also combined with the real training data specified in Table 2, with no synthesized data added to the validation set. Table 1 describes all the datasets that were generated and their contents.

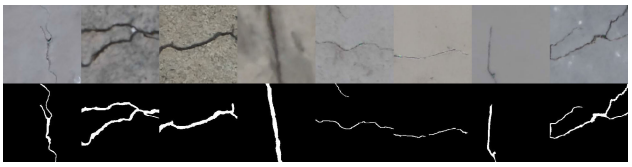
By training models with datasets augmented with motion, zoom, and defocus blurs, we significantly bolster their ability to interpret and analyze crack images under various real-world conditions. This approach not only enhances the model's generalization capabilities but also ensures its practical applicability in diverse structural health monitoring scenarios. The enhanced model is better equipped to handle variations in image clarity and texture, which are common in real-world structural assessments, thereby improving its reliability and effectiveness in practical applications. Incorporating zoom, defocus, and motion blur augmentations into our training datasets prepares the DeepLabV3+ model for adverse camera conditions often encountered in practical settings. Later, we introduce ensemble modeling to leverage the different synthesized augmented datasets to increase MeanIoU performance.

B. HYPERPARAMETER TUNING DEEPLABV3+

1) DEEPLABV3+

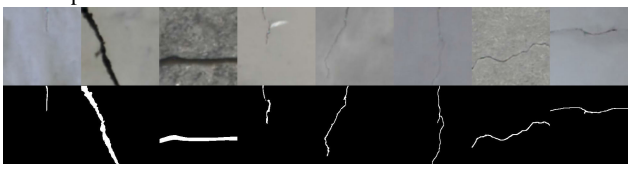
DeepLabV3+[2] is a state-of-the-art model for semantic segmentation known for its effectiveness in handling complex image segmentation tasks. The architecture of DeepLabV3+ is an evolution of its predecessors, designed to provide

[2]DeepLabV3+ implementation and pretrained model can be accessed from this GitHub repository: https://github.com/VainF/DeepLabV3Plus-Pytorch/tree/master

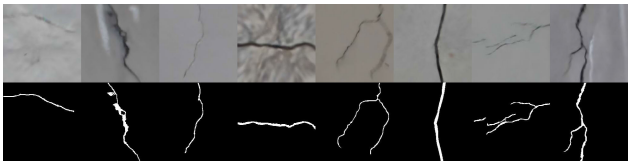


(a) Original Pixel Space BBDM Sample Images: After 600,000 steps of training, the images exhibit the diversity of domains when reproducing crack annotations. The pixel space images very accurately represent the annotation images when compared to the latent space.

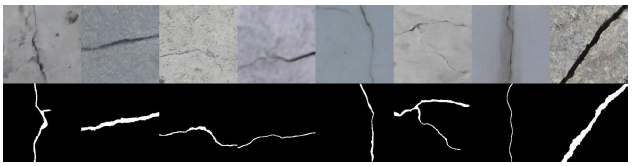
(b) Motion Blur Pixel Space BBDM Sample Images: Motion blur was applied to all real synthesized datasets with a random severity of 3 to 12.



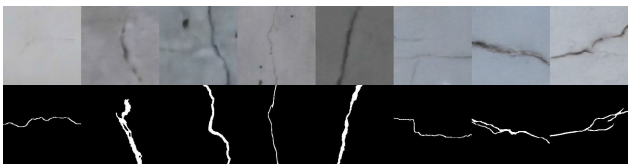
(c) Defocus Blur Pixel Space BBDM Sample Images: Defocus emulates a camera that is not focused when the image is captured. A severity magnitude of one was applied to all images.



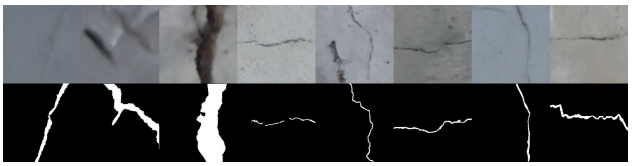
(d) Zoom Blur Pixel Space BBDM Sample Images: A severity magnitude of one was applied to all synthesized images.



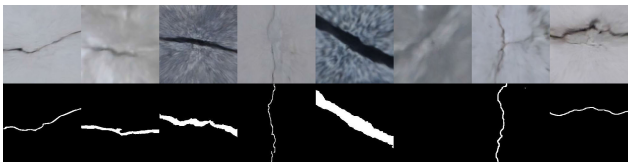
(e) Original Latent Space BBDM Sample Images: The LBBDM was trained for efficiency and generalization, producing images with diverse and realistic backgrounds when compared to the pixel space. The latent space synthesized real images do not adhere as closely to the annotation images as the pixel space.



(f) Motion Blur Latent Space BBDM Sample Images: A random severity in the range of 3 to 12 was applied to the latent space images.



(g) Defocus Blur Latent Space BBDM Sample Images: A strength of 1 blur was applied to the synthesized images.



(h) Zoom Blur Latent Space BBDM Sample Images: A strength of 1 was applied to the synthesized images.

FIGURE 4. **Sample images from synthesized BBDM datasets.**

high precision in object boundary delineation. DeepLabV3+ employs atrous (dilated) convolutions, allowing the model to control the resolution at which feature responses are computed within Deep Convolutional Neural Networks. DeepLabV3+ was chosen for this feature because it is particularly beneficial for segmenting fine details in images, such as small or narrow cracks. At the heart of DeepLabV3+ is the ASPP module, which applies atrous convolution at multiple rates to capture multi-scale information. The module allows for detecting cracks that vary in scale. Additionally, the model adopts an encoder-decoder structure. The encoder module, augmented with ASPP, captures rich contextual information, while the decoder module refines the segmentation results, especially along object boundaries. This refinement is vital for the precise delineation of crack contours.

Within the framework of our study, the employment of DeepLabV3+ extends well beyond its utilization as a mere advanced tool for semantic segmentation. A dedicated effort was undertaken to conduct extensive hyperparameter tuning, adapting the DeepLabV3+ architecture to meet the specific demands of our dataset and the intricacies of our segmentation objectives. Moreover, DeepLabV3+ was instrumental as a benchmarking model to assess the efficacy of our innovative data generation methodologies, which encompass the synthesis of training data via StyleGAN and BBDM. Additionally, we delve into the realm of ensemble modeling utilizing DeepLabV3+, aiming to bolster segmentation performance. By orchestrating the training of multiple DeepLabV3+ models across a spectrum of synthesized datasets and subsequently amalgamating their outputs, we witnessed a notable enhancement in segmentation

TABLE 1. Description of training datasets.

Dataset	Description
BCL	Training and Validation dataset split described in 2 (BCL Training and Val).
BCL + BCL 2.0	Combines BCL Training and Val with data from BCL 2.0 images. BCL 2.0 was produced by Jin et al [3] which was trained on the original BCL dataset and contains fully synthesized images for segmentation.
M-BBDM-P	7,800 BBDM pixel space synthesized images were combined with BCL Training and Val.
M-BBDM-P-ZB	M-BBDM-P with zoom blur applied to synthesized images.
M-BBDM-P-MB	M-BBDM-P with motion blur at severity levels 3 to 12 to only synthesized images.
M-BBDM-P-DB	M-BBDM-P with defocus blur applied to only synthesized images.
M-BBDM-L	7,800 BBDM latent space synthesized images were combined with BCL training and validation.
M-BBDM-L-ZB	M-BBDM-L with zoom blur applied to only synthesized images.
M-BBDM-L-MB	M-BBDM-L with motion blur at severity levels 3 to 12 was applied to only synthesized images.
M-BBDM-L-DB	M-BBDM-L with defocus blur applied to only synthesized images.

precision. DeepLabV3+ is the backbone for evaluating our segmentation performance which is why hyperparameter tuning is paramount.

2) DEEPLABV3+ HYPERPARAMETER TUNING

The training of DeepLabV3+ was meticulously conducted on a dataset that encompasses a wide spectrum of real-world conditions. The composition of the datasets used for training, validation, and testing is detailed in Table 2.

Our primary objective in hyperparameter tuning was to refine the DeepLabV3+ model for optimal crack segmentation performance. We began our experiments using a basic setup of the DeepLabV3+ model. This initial setup involved training the model for 200 epochs. We set the learning rate at 0.001 and used a batch size of 128. The model's output stride was fixed at 8. For optimization, we chose the Adam optimizer, applying a weight decay of 0.0001. The optimizer's beta values were set at 0.9 and 0.999. We utilized a cross-entropy loss function, and the backbone of the model was MobileNet. This initial configuration yielded an accuracy of 59.57% (results summarized in Table 3).

After extensive experimentation, we identified the optimal hyperparameters for the DeepLabV3+ model, detailed in Table 3, resulting in the best following results: the model was trained over 200 epochs, utilizing a fixed learning rate of 0.001 and a batch size of 128. We maintained an output stride of 8 and employed the Adam optimizer, featuring a weight decay of 0.0001 and betas set at 0.9 and 0.999 throughout the training process. To facilitate model convergence and performance, we applied a cross-entropy loss function. Our chosen backbone architecture was based on a pre-trained 32-layer High-Resolution Net, and we further enhanced training robustness by implementing on-the-fly data augmentation.

TABLE 2. Composition of the training, validation, and test datasets for hyperparameter tuning.

Dataset Group	Image Types	Number of Images
Training	Cracks	4,036
	Noise	3,195
	Steel	1,424
	Synthetic - BCL 2.0	10,919
Validation	Cracks	866
	Noise	0
	Steel	305
	Synthetic - BCL 2.0	2,340
Test	Cracks	866
	Steel	306

C. ENSEMBLE MODELING DEEPLABV3+ MODELS TRAINED ON VARIOUS SYNTHESIZED DATASETS

To enhance the performance of structural crack detection, we ensemble multiple DeepLabV3+ models, each trained on the various datasets mentioned in Table 1. For each of the six augmented datasets, a DeepLabV3+ model was independently trained. This approach allowed each model to specialize in images with particular characteristics of blur, thereby enhancing its detection capabilities. Upon training completion, these models were ensembled by pixel or latent space using majority voting to evaluate their collective performance on the testing set outlined in Table 2. Majority voting is an effective technique for consolidating outputs from multiple DeepLabV3+ models. This method involves each model in the ensemble independently classifying each pixel in an image as a crack or not. For every pixel, the classification that receives the majority vote across all models determines the final prediction. This approach enhances the accuracy of the predictions by mitigating individual model biases and errors, which is especially beneficial in handling the diverse image qualities inherent in crack detection tasks. Refer to Figure 5 for sample results of majority voting ensemble modeling.

IV. EXPERIMENTS

A. BRIDGE CRACK LIBRARY DATASET

The Bridge Crack Library (BCL) dataset, developed by Ye et al. [62] in their seminal work on structural crack detection using pruned fully convolutional networks, represents a significant advancement in the domain of automated crack detection for in-service bridges. This comprehensive dataset comprises 11,000 pixel-wise labeled images of 256 by 256 resolution that were meticulously curated to include a diverse range of crack forms across various structural materials, including masonry, concrete, and steel. Notably, the dataset encapsulates 5,769 nonsteel crack images, 2,036 steel crack images, and 3,195 images categorized as noise, which were derived from the examination of over 50 bridges by experienced inspection teams over two years. Refer to back to Figure 2 The meticulous data collection process undertaken for the BCL dataset underscores its value in the field of structural engineering. By employing multiple cameras and covering a wide array of in-service bridges,

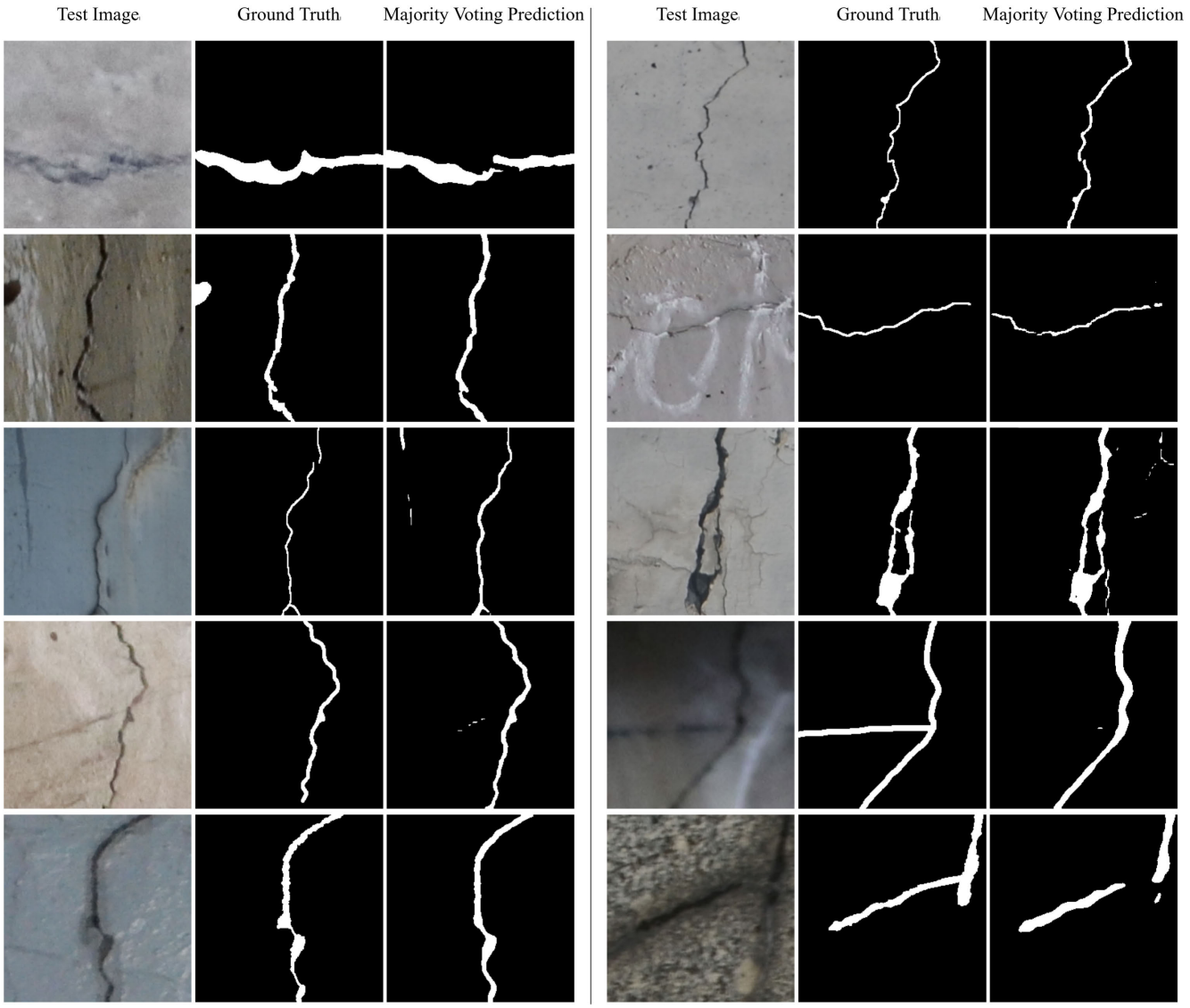


FIGURE 5. Majority Voting Ensemble Predictions. This majority voting ensemble method effectively combines individual model predictions of three DeepLabV3+ models that were trained separately on M-BBDM-P-ZB, M-BBDM-P-DB, and M-BBDM-P-MB.

the dataset captures an extensive range of crack images and noise motifs, set against different environmental conditions and structural backgrounds.

A pivotal aspect of the BCL dataset's creation was the pixel-wise annotation of images, a labor-intensive process that ensures high precision in crack delineation. The manual annotation process was supported by specialized software that facilitated precise labeling. Annotators used digital pens and tablets, which offered more accuracy and control than traditional mouse-based interfaces, allowing for the detailed tracing of crack contours. This hands-on approach ensured that the dataset captured the nuances of each crack, including its width, direction, and branching patterns.

The BCL dataset's comprehensive and precisely annotated images provide a robust foundation for generating synthetic datasets. These synthetic datasets, in turn, are instrumental in augmenting the diversity of training samples available for DNNs, thus addressing one of the primary challenges in the application of deep learning for crack detection: the scarcity of labeled training data.

B. DEEPLABV3+ HYPERPARAMETER TUNING RESULTS

In our experiments, we iteratively introduced improvements to the DeepLabV3+ model and applied these enhancements cumulatively in subsequent tests. This approach allowed us to systematically assess the impact of each modification on the MeanIoU performance. Table 3 below provides a summary of these changes and their respective impacts.

The experiments reveal a significant influence of the backbone architecture on the model's performance. While the baseline model with MobileNet as the backbone achieved respectable accuracy, switching to the ResNet-50 and HRNet architectures led to noticeable improvements. Notably, the HRNetv2 with 32 layers stood out as the most effective

TABLE 3. Summary of hyperparameter tuning for DeepLabV3+.

Improvements	Loss Function	Learning Rate	Augmentation	Backbone	Batch Size	MeanIoU (%)	Delta (%)
Jin et al [3]: BCL + BCL 2.0	Cross-Entropy Loss	0.001	No	-	128	54.23	-
MILAB Improvements							
Baseline: Cross-Entropy Loss	Cross-Entropy	0.001	No	MobileNet	128	59.57	+5.34
Focal Loss	Focal Loss	0.001	No	MobileNet	128	59.93	+5.70
New Learning Rate	Cross-Entropy Loss	0.0003	No	MobileNet	128	58.86	+4.63
On the Fly Data Augmentation	Cross-Entropy Loss	0.001	Yes	MobileNet	128	62.44	+8.21
Aug + Dice Loss Variants	Dice	0.001	Yes	MobileNet	128	59.98	+5.75
Aug + Focal Loss Variants	Focal	0.001	Yes	MobileNet	128	61.57	+7.34
ResNet-50 Backbone	Cross-Entropy Loss	0.001	Yes	ResNet-50	128	62.12	+7.89
HRNetv2-48	Cross-Entropy Loss	0.001	Yes	HRNetv2-48	64	60.52	+6.29
HRNetv2-32, Pretrained	**Cross-Entropy Loss**	**0.001**	**Yes**	**HRNetv2-32**	**128**	**63.43**	**+9.20**

TABLE 4. Testing results for synthesized datasets.

Model/Dataset	MeanIoU (%)	Precision (%)	Recall (%)	Accuracy (%)	F1 Score (%)
Ye et al. [15]	47.57	81.03	53.53	98.64	64.47
Jin et al. [3]	54.23	62.35	77.10	98.72	68.91
BCL	63.30	68.46	78.33	98.32	70.60
BCL + BCL 2.0	63.43	73.16	70.94	98.46	69.38
M-BBDM-P	61.36	69.11	74.37	98.29	68.88
M-BBDM-P-ZB	64.75	70.59	75.44	98.44	70.59
M-BBDM-P-MB	63.35	72.95	70.37	98.46	69.02
M-BBDM-P-DB	64.14	67.28	79.88	98.36	70.82
M-BBDM-P-E-Voting	65.62	71.20	75.26	98.51	71.04
M-BBDM-P-E-Mean	65.33	71.39	75.48	98.50	71.14
M-BBDM-P-E-Or	63.89	66.01	83.20	98.27	71.79
M-BBDM-L	62.22	66.33	75.47	98.26	67.79
M-BBDM-L-ZB	63.82	72.79	70.59	98.47	69.17
M-BBDM-L-MB	63.43	72.61	70.46	98.48	68.85
M-BBDM-L-DB	64.01	73.51	69.64	98.48	68.77
M-BBDM-L-E-Voting	64.40	73.42	70.31	98.50	69.22
M-BBDM-L-E-Mean	64.38	73.28	70.23	98.51	69.16
M-BBDM-L-E-Or	64.61	70.90	75.95	98.43	71.01

backbone, achieving the highest accuracy of 63.43%. This indicates that the HRNetv2 architecture, particularly with 32 layers, is better suited for capturing the detailed features necessary for crack segmentation. For the loss functions, we tried focal loss and dice loss but found that the basic cross-entropy loss function consistently gave us good results in terms of accuracy. The learning rate, while a critical learning rate in many deep learning applications, did not significantly alter our outcomes. We opted to keep learning rate tests to a minimum to focus on identifying other hyperparameters that contribute more substantially to accuracy improvements. The introduction of on-the-fly data augmentation marked a significant improvement in model performance, boosting accuracy to 62.43%. The augmentation techniques, including random scale, crop, horizontal and vertical flips, random rotation, and color jitter, contributed to the model's improved generalization and robustness against varying crack patterns and lighting conditions. The experiment conducted by Jin et al. [3] using BCL plus BCL 2.0 datasets yielded a lower accuracy of 54.23%, highlighting the effectiveness of the hyperparameter tuning and modifications applied in our approach.

C. USING HYPERTUNED DEEPLABV3+ FOR SEMANTIC SEGMENTATION OF GENERATED DATASETS

These results, in Table 4, highlight the capabilities of the BBDM in synthesizing realistic crack images and also demonstrate the effectiveness of various data augmentation and ensemble modeling techniques in improving MeanIoU. Additionally, we use several segmentation metrics to assess the performance of the model's accuracy and reliability.

- **Mean Intersection over Union (MeanIoU):** MeanIoU is especially relevant for segmentation tasks because it calculates the intersection over union for each class and then averages these values.
- **Precision:** Precision is defined as the ratio of correctly predicted positive observations to the total predicted positives. In the context of crack segmentation, it reflects the accuracy of the model in identifying true cracks as opposed to false positives. A higher precision indicates

that the model is more effective in correctly labeling crack pixels, minimizing the instances where non-crack pixels are incorrectly classified as cracks.

- **Recall:** Recall measures the ratio of correctly predicted positive observations to the total actual positives. Recall score quantifies the model's ability to identify all actual crack instances in the images. High recall is crucial in applications like crack detection, as it ensures that the model captures as many true crack occurrences as possible, reducing the risk of missing critical defects. However, higher recall typically comes with the cost of an increased number of false positives.
- **Accuracy:** This metric is the ratio of correctly predicted observations (both true positives and true negatives) to the total observations in the dataset. While accuracy is a widely used metric, it does not provide a comprehensive assessment of performance, especially in the cases of imbalanced datasets where one class (e.g., cracks) is significantly underrepresented when compared to another (e.g., non-cracks).
- **F1-Score:** The F1-Score is the harmonic mean of precision and recall. It provides a balanced measure between these two metrics, making it particularly suitable for situations where there is an uneven class distribution, as is often the case in crack detection datasets.

The M-BBDM-P datasets, particularly the ones with zoom blur and defocus blur showed high MeanIoU scores, with zoom blur leading at 64.75% and defocus blur following at 64.14%. Furthermore, each augmented dataset exhibited distinct strengths: the pixel space datasets generally outperformed in MeanIoU, Recall, and F1-score, whereas the latent space datasets excelled in precision and accuracy. All augmented datasets demonstrated an improvement in MeanIoU compared to their predecessors that were not augmented. The pixel space saw at least a 2% increase in MeanIoU performance, while the latent space saw at least a 1.21% increase.

Since the accuracy between the augmented models in each of their respective spaces was similar, we tried out three different ensemble modeling techniques: majority voting, mean, and logical or. Mean ensemble modeling was applied by averaging the predictions of all models, which helped in smoothing out the anomalies and uncertainties in individual predictions. Logically OR'ing prediction images involves combining the outputs of multiple models such that if any model predicts a pixel as a crack, it is marked as a crack in the final aggregated image, enhancing the detection sensitivity. Mostly, individual models may have unique strengths and weaknesses; by combining them, the ensemble can leverage the strengths of each while mitigating their weaknesses.

Ensemble modeling proved to almost always improve MeanIoU performance and F1-score. Specifically, majority voting performed the best amongst both the pixel and latent space, achieving 65.62% and 64.40% MeanIoU respectively. However, mean and majority voting ensemble modeling both performed very similarly across all performance metrics. Logically or ensemble modeling tended to score the best in terms of recall and F1-score when compared to other techniques likely because of its increased sensitivity. Illustrated in the Appendix in Figure 6, the final results of the ensemble modeling methods can be found.

The M-BBDM-L-DB scores a competitive precision score, of 73.51%. The M-BBDM-P-Ensemble-Voting model, with the highest Recall (83.20%), and F1 Score (71.79%), proved its efficacy in identifying cracks more comprehensively than the baseline models. The M-BBDM-P-Ensemble-Voting model exhibited the highest MeanIoU, surpassing the baselines and scoring 65.62%.

1) PROPOSED METHOD DISCUSSION

For any practical application, we need to consider a combination of metrics that align closely with the specific requirements of its use case. Selecting the most suitable method requires a nuanced understanding of the trade-offs inherent in different performance metrics.

Given the nature of our application, where missing a crack could have serious consequences, a high recall rate is essential. Thus, we aim to capture as many cracks as possible, even at the risk of some false positives. M-BBDM-P-E-Or achieved the highest recall score of 83.20%.

If the primary goal is to provide a balanced evaluation between precision and recall then F1 Score emerges as an apt metric. The F1 Score, the harmonic mean of precision and recall, offers a comprehensive assessment of model performance. Therefore, model M-BBDM-P-E-Or also emerges as the best solution under this criteria. This model demonstrates an effective balance, achieving a high recall without significantly compromising on precision, as evident in its F1-Score (71.79%).

If the priority shifts towards minimizing false positives (increase precision), while still maintaining a high F1 score, there are better-suited models for this task. In this case, models employing mean and majority voting ensemble techniques, specifically M-BBDM-P-E-Mean and M-BBDM-P-E-Voting, become more relevant because of their high F1 score and high MeanIoU and precision.

In light of both scenarios, the M-BBDM-P-E-Or model distinguishes itself as the superior choice, boasting an exceptionally high detection rate coupled with an impressive F1 score.

V. CONCLUSION

Our study made significant contributions to the field of structural crack detection, particularly in demonstrating the efficacy of ensemble models trained on augmented synthetic datasets to enhance semantic segmentation performance. By employing StyleGAN3 and BBDM for data synthesis, we created rich and varied datasets that significantly improved the training of DeepLabV3+ models on structural health datasets. This advancement addresses a critical challenge in the field - the scarcity of diverse and high-quality datasets for training sophisticated segmentation

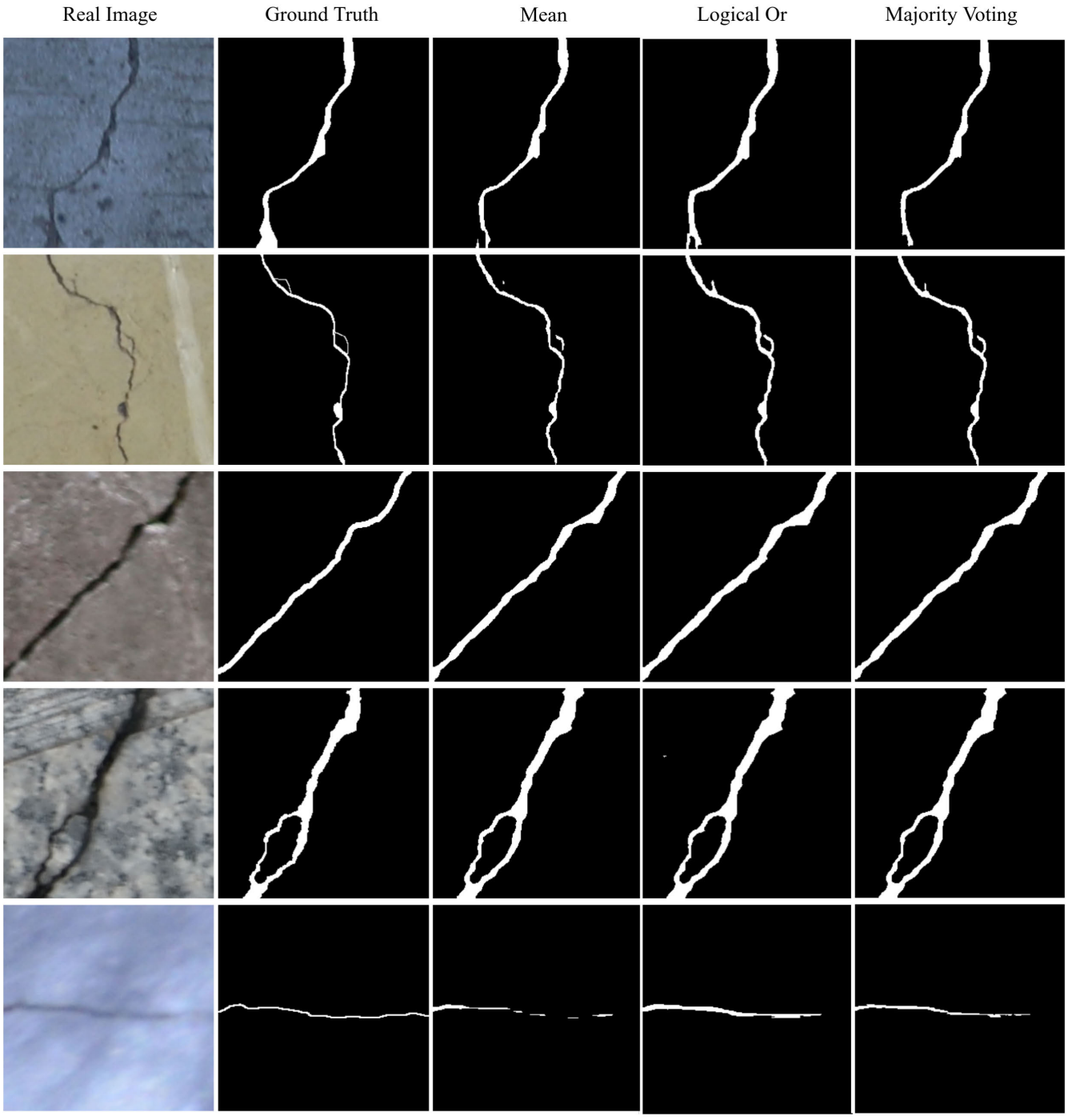


FIGURE 6. Final prediction results from ensemble modeling synthesized BBDM datasets.

models. A pivotal contribution of our research lies in the meticulous hyperparameter tuning of the DeepLabV3+ model. Our findings underscore the importance of hyperparameter optimization, revealing how subtle changes in model configuration can lead to substantial improvements in accuracy and efficiency. By experimenting with different backbones, optimizers, and on-the-fly data augmentation techniques, we achieved marked improvements in crack segmentation accuracy. The selection of HRNetv2 with 32 layers as the backbone, in particular, led to the highest accuracy of 63.43%, highlighting the substantial impact of model architecture on segmentation performance. One of the novel aspects of our research was the application of diffusion models, specifically BBDM, to generate synthetic datasets. We showed that these models could create realistic and diverse crack images, which, when used in conjunction with real data, lead to improved segmentation performance when combined with ensemble modeling methods and blur augmentation techniques. This finding opens new possibilities in data generation for machine learning, extending beyond the

realm of crack detection. Looking ahead, the methodologies and insights gleaned from our work hold promise for broader application. Future work could explore the applicability of our ensemble modeling approach and data synthesis techniques to other crack detection datasets and within different domains, such as aerospace, civil infrastructure, and manufacturing. By training on more different types of domains with crack-like defects, we can explore the capabilities of automated inspection systems across a wider range of applications. Ultimately, by extending our research into these new territories, we aim to further enhance the capabilities of machine learning models in identifying and quantifying structural damage, thereby contributing to the safety and longevity of critical infrastructure worldwide.

APPENDIX

See the Figure 6.

ACKNOWLEDGMENT

The authors extend their gratitude to the MILAB / Graduate Student Victor Philippe, for this support and whose insights greatly contributed to this work. Additionally, Thirimachos Bourlai would like to recognize the assistance of ChatGPT [66], developed by OpenAI, for the editing process. The efforts of the OpenAI team in creating and making this language model accessible are duly appreciated by all authors.

REFERENCES

[1] S. K. U. Rehman, Z. Ibrahim, S. A. Memon, and M. Jameel, "Non-destructive test methods for concrete bridges: A review," *Construct. Building Mater.*, vol. 107, pp. 58–86, Mar. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950061815306905

[2] R. Ali, J. H. Chuah, M. S. A. Talip, N. Mokhtar, and M. A. Shoaib, "Structural crack detection using deep convolutional neural networks," *Autom. Construct.*, vol. 133, Jan. 2022, Art. no. 103989.

[3] T. Jin, X. W. Ye, and Z. X. Li, "Establishment and evaluation of conditional GAN-based image dataset for semantic segmentation of structural cracks," *Eng. Struct.*, vol. 285, Jun. 2023, Art. no. 116058.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[5] T. Jin, Z. Li, Y. Ding, S. Ma, and Y. Ou, "Bridge crack library," Harvard Dataverse, 2021, doi: 10.7910/DVN/RURXSH.

[6] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[7] H. Fu, D. Meng, W. Li, and Y. Wang, "Bridge crack semantic segmentation based on improved DeepLabV3+," *J. Mar. Sci. Eng.*, vol. 9, no. 6, p. 671, Jun. 2021.

[8] J. Wang, Y. Liu, X. Nie, and Y. L. Mo, "Deep convolutional neural networks for semantic segmentation of cracks," *Struct. Control Health Monitor.*, vol. 29, no. 1, p. e2850, Jan. 2022.

[9] S. O. Atik, M. E. Atik, and C. Ipbuker, "Comparative research on different backbone architectures of DeepLabV3+ for building segmentation," *J. Appl. Remote Sens.*, vol. 16, no. 2, May 2022, Art. no. 024510.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[11] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[13] Z. Nie, J. Xu, and S. Zhang, "Analysis on DeepLabV3+ performance for automatic steel defects detection," 2020, *arXiv:2004.04822*.

[14] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[15] X. Ye, Z. Li, and T. Jin, "Smartphone-based structural crack detection using pruned fully convolutional networks and edge computing," *Smart Struct. Syst.*, vol. 29, no. 1, pp. 141–151, 2022.

[16] Z. Fan, C. Li, Y. Chen, P. D. Mascio, X. Chen, G. Zhu, and G. Loprencipe, "Ensemble of deep convolutional neural networks for automatic pavement crack detection and measurement," *Coatings*, vol. 10, no. 2, p. 152, Feb. 2020.

[17] D. Hirata and N. Takahashi, "Ensemble learning in CNN augmented with fully connected subnetworks," *IEICE Trans. Inf. Syst.*, vol. 106, no. 7, pp. 1258–1261, 2023.

[18] V. Kailkhura, S. Aravindh, S. S. Jha, and N. Jayanthi, "Ensemble learning-based approach for crack detection using CNN," in *Proc. 4th Int. Conf. Trends Electron. Informat.*, Jun. 2020, pp. 808–815.

[19] F. J. Rodriguez-Lozano, F. León-García, J. C. Gámez-Granados, J. M. Palomares, and J. Olivares, "Benefits of ensemble models in road pavement cracking classification," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 35, no. 11, pp. 1194–1208, Nov. 2020.

[20] G. Li, B. Ma, S. He, X. Ren, and Q. Liu, "Automatic tunnel crack detection based on U-Net and a convolutional neural network with alternately updated clique," *Sensors*, vol. 20, no. 3, p. 717, Jan. 2020.

[21] A. A. Maarouf and F. Hachouf, "Transfer learning-based ensemble deep learning for road cracks detection," in *Proc. Int. Conf. Adv. Aspects Softw. Eng. (ICAASE)*, Sep. 2022, pp. 1–6.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 84–90.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[26] J. Tang, C. Chen, Z. Huang, X. Zhang, W. Li, M. Huang, and L. Deng, "Crack Unet: Crack recognition algorithm based on three-dimensional ground penetrating radar images," *Sensors*, vol. 22, no. 23, p. 9366, Dec. 2022.

[27] C. Han, T. Ma, J. Huyan, X. Huang, and Y. Zhang, "CrackW-Net: A novel pavement crack image segmentation convolutional neural network," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 22135–22144, Nov. 2022.

[28] J. Liu, X. Yang, S. Lau, X. Wang, S. Luo, V. C. Lee, and L. Ding, "Automated pavement crack detection and segmentation based on two-step convolutional neural network," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 35, no. 11, pp. 1291–1305, Nov. 2020.

[29] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[30] S. L. H. Lau, E. K. P. Chong, X. Yang, and X. Wang, "Automated pavement crack segmentation using U-Net-based convolutional neural network," *IEEE Access*, vol. 8, pp. 114892–114899, 2020.

[31] W. Wang and C. Su, "Convolutional neural network-based pavement crack segmentation using pyramid attention network," *IEEE Access*, vol. 8, pp. 206548–206558, 2020.

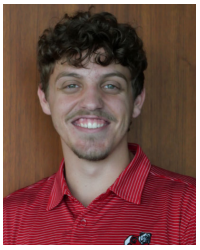
[32] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[33] J. Huyan, W. Li, S. Tighe, Z. Xu, and J. Zhai, "CrackU-Net: A novel deep convolutional neural network for pixelwise pavement crack detection," *Struct. Control Health Monitor.*, vol. 27, no. 8, p. e2551, Aug. 2020.

[34] W. Choi and Y.-J. Cha, "SDDNet: Real-time crack segmentation," *IEEE Trans. Ind. Electron.*, vol. 67, no. 9, pp. 8016–8025, Sep. 2020.

[35] G. Yu, J. Dong, Y. Wang, and X. Zhou, "RUC-Net: A residual-Unet-based convolutional neural network for pixel-level pavement crack segmentation," *Sensors*, vol. 23, no. 1, p. 53, Dec. 2022.

[36] S. Wang and W. Tang, "Pavement crack segmentation algorithm based on local optimal threshold of cracks density distribution," in *Proc. Int. Conf. Intell. Comput.*, Zhengzhou, China. Cham, Switzerland: Springer, 2011, pp. 298–302.
[37] L. Cui, Z. Qi, Z. Chen, F. Meng, and Y. Shi, "Pavement distress detection using random decision forests," in *Proc. Int. Conf. Data Sci.* Cham, Switzerland: Springer, 2015, pp. 95–102.
[38] F. Yang, L. Zhang, S. Yu, D. Prokhorov, X. Mei, and H. Ling, "Feature pyramid and hierarchical boosting network for pavement crack detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 4, pp. 1525–1535, Apr. 2020.
[39] V. Polovnikov, D. Alekseev, I. Vinogradov, and G. V. Lashkia, "DAUNet: Deep augmented neural network for pavement crack segmentation," *IEEE Access*, vol. 9, pp. 125714–125723, 2021.
[40] Z. Qu, C. Cao, L. Liu, and D.-Y. Zhou, "A deeply supervised convolutional neural network for pavement crack detection with multiscale feature fusion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4890–4899, Sep. 2022.
[41] Z. Qu, Y. Li, and Q. Zhou, "CrackT-Net: A method of convolutional neural network and transformer for crack segmentation," *J. Electron. Imag.*, vol. 31, no. 2, Apr. 2022, Art. no. 023040.
[42] Z. Wang, J. Yang, H. Jiang, and X. Fan, "CNN training with twenty samples for crack detection via data augmentation," *Sensors*, vol. 20, no. 17, p. 4849, Aug. 2020.
[43] L. Zhang, F. Yang, Y. Daniel Zhang, and Y. J. Zhu, "Road crack detection using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3708–3712.
[44] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 12, pp. 3434–3445, Dec. 2016.
[45] Y. Liu, J. Yao, X. Lu, R. Xie, and L. Li, "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation," *Neurocomputing*, vol. 338, pp. 139–153, Apr. 2019.
[46] Z. Pan, S. L. H. Lau, X. Yang, N. Guo, and X. Wang, "Automatic pavement crack segmentation using a generative adversarial network (GAN)-based convolutional neural network," *Results Eng.*, vol. 19, Sep. 2023, Art. no. 101267.
[47] X. Yu, G. Li, W. Lou, S. Liu, X. Wan, Y. Chen, and H. Li, "Diffusion-based data augmentation for nuclei image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2023, pp. 592–602.
[48] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 8780–8794.
[49] K. Ding, M. Zhou, H. Wang, O. Gevaert, D. Metaxas, and S. Zhang, "A large-scale synthetic pathological dataset for deep learning-enabled segmentation of breast cancer," *Sci. Data*, vol. 10, no. 1, p. 231, Apr. 2023.
[50] L. Fetty, M. Bylund, P. Kuess, G. Heilemann, T. Nyholm, D. Georg, and T. Löfstedt, "Latent space manipulation for high-resolution medical image synthesis via the StyleGAN," *Zeitschrift Medizinische Physik*, vol. 30, no. 4, pp. 305–314, Nov. 2020.
[51] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, and N. Rajpoot, "Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101563.
[52] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 852–863.
[53] J. Xu, C. Yuan, J. Gu, J. Liu, J. An, and Q. Kong, "Innovative synthetic data augmentation for dam crack detection, segmentation, and quantification," *Struct. Health Monitor.*, vol. 22, no. 4, pp. 2402–2426, Jul. 2023.
[54] C. Bartz, H. Raetz, J. Otholt, C. Meinel, and H. Yang, "Synthesis in style: Semantic segmentation of historical documents using synthetic data," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 3878–3884.
[55] M. Boillet, C. Kermorvant, and T. Paquet, "Multiple document datasets pre-training improves text line detection with deep neural networks," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 2134–2141.
[56] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, "Expectation-maximization attention networks for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9166–9175.
[57] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
[58] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021, *arXiv:2112.10752*.
[59] B. Li, K. Xue, B. Liu, and Y.-K. Lai, "BBDM: Image-to-image translation with Brownian bridge diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1952–1961.
[60] G. Zhai, Y. Narazaki, S. Wang, S. A. V. Shajihan, and B. F. Spencer Jr., "Synthetic data augmentation for pixel-wise steel fatigue crack identification using fully convolutional networks," *Smart Struct. Syst.*, vol. 29, no. 1, pp. 237–250, 2022.
[61] L. Pei, Z. Sun, L. Xiao, W. Li, J. Sun, and H. Zhang, "Virtual generation of pavement crack images based on improved deep convolutional generative adversarial network," *Eng. Appl. Artif. Intell.*, vol. 104, Sep. 2021, Art. no. 104376.
[62] X. W. Ye, T. Jin, Z. X. Li, S. Y. Ma, Y. Ding, and Y. H. Ou, "Structural crack detection from benchmark data sets using pruned fully convolutional networks," *J. Struct. Eng.*, vol. 147, no. 11, Nov. 2021, Art. no. 04721008.
[63] T. Jin and Z. Li, "Bridge crack library 2.0," Harvard Dataverse, 2022, doi: 10.7910/DVN/TUFAJT.
[64] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[65] C. Michaelis, B. Mitzkus, R. Geirhos, E. Rusak, O. Bringmann, A. S. Ecker, M. Bethge, and W. Brendel, "Benchmarking robustness in object detection: Autonomous driving when winter is coming," 2019, *arXiv:1907.07484*.
[66] OpenAI. (2023). *ChatGPT: Optimizing Language Models for Dialogue*. Accessed: Jan. 2, 2024. [Online]. Available: https://www.openai.com/blog/chatgpt



COLLINS C. RAKOWSKI was born in Atlanta, GA, USA, in 2000. He received the degree in computer systems engineering from the University of Georgia, Athens, GA, USA, in 2023. He is currently pursuing the master's degree in engineering, with a focus on artificial intelligence. He was a Software Engineer with NCR, Atlanta, and a Building Facility Assessment Engineer with WOOD, Atlanta. Becoming a Software Engineer helped guide me into the world of machine learning and deep learning. He is continuing his studies with the University of Georgia, while actively engaging in research projects that explore advancements in hyperspectral imaging. His major field of study was computer systems. He has gained real-world experience through internships. His dedication to his field is also evident in his commitment to ongoing professional development and contributions to various technical committees.

THIRIMACHOS BOURLAI was an Adjunct Faculty with the Lane Department of Computer Science and Engineering and the School of Medicine, West Virginia University (WVU). He is currently an Associate Professor with the School of Electrical and Computer Engineering, University of Georgia. In addition to his academic role, he holds a Joint Appointment with Savannah River National Laboratories. He is the Founder and the Director of the Multi-Spectral Imagery Laboratory and holds significant editorial roles. Complementing his innovative work are three patents and a prolific publication record, encompassing more than 140 contributions to journals, conferences, book chapters, and magazines in the domains of computer vision, biometrics, and related fields. His substantial research contributions are evident in his authorship of four notable books with Springer, including *Face Recognition Across the Imaging Spectrum* (2016), *Surveillance in Action* (2018), *Securing Social Identity in Mobile Platforms* (2020), and *Disease Control Through Social Network Surveillance* (2022). In the realm of organizational leadership, he is a member of the Board of Directors of the Document Security Alliance and the Academic Research and Innovation Expert Group of the Biometrics Institute. He is a Series Editor of *Advanced Sciences and Technologies for Security Applications* and an Associate Editor of *Pattern Recognition* (Elsevier) and *IET Electronics Letters*. He is the Former VP of Education of the IEEE Biometrics Council.