Enhancing Deep Learning-Based Damage Segmentation with Depth Hallucination

TARUTAL GHOSH MONDAL and MOHAMMAD R. JAHANSHAHI

ABSTRACT

Recent studies have suggested that the fusion of cross-modal information can enhance the performance of deep learning-based segmentation algorithms. In this context, this study evaluates the benefits of RGB-D fusion with regard to damage segmentation in reinforced concrete buildings. The fusion of depth data was observed to enhance the segmentation performance significantly. Additionally, a number of surrogate techniques based on modality hallucination and monocular depth estimation are exploited to eliminate the need for depth sensing at test time without foregoing the benefits of depth fusion. The proposed techniques require depth data only for network training, and at test time, depth features are simulated from the corresponding RGB frames, obliterating the need for real depth perception. The proposed methods are evaluated and are shown to increase the damage segmentation accuracy.

INTRODUCTION

Aging civil infrastructures require periodic inspection in order to prevent sudden failure, which causes loss of lives and economic setbacks. The existing inspection techniques are, by and large, manual and, therefore, time-consuming, subjective, expensive, and risky. Computer vision-based algorithms have been explored in recent times to investigate the prospect of robotic inspection as a viable alternative to such manual techniques. A number of studies exploited deep learning-based methods to this end for autonomous defect detection in civil infrastructures [1, 2]. However, the previous studies relied solely on the photometric (RGB) data for identifying damages in videos and images. There is no study to date that leveraged depth perception for semantic labeling of various damages that commonly occur in reinforced concrete structures subjected to extreme loading. The present study addresses this research gap by incorporating depth fusion into an encoder-decoder-based fully convolutional network for semantic damage

Tarutal Ghosh Mondal, Post-doctoral Fellow, Email: tg5qf@mst.edu. Department of Civil, Architectural and Environmental Engineering, Missouri University of Science and Technology, Rolla, MO, USA. Mohammad R. Jahanshahi, Associate Professor, Email: jahansha@purdue.edu. Lyles School of Civil Engineering, Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA

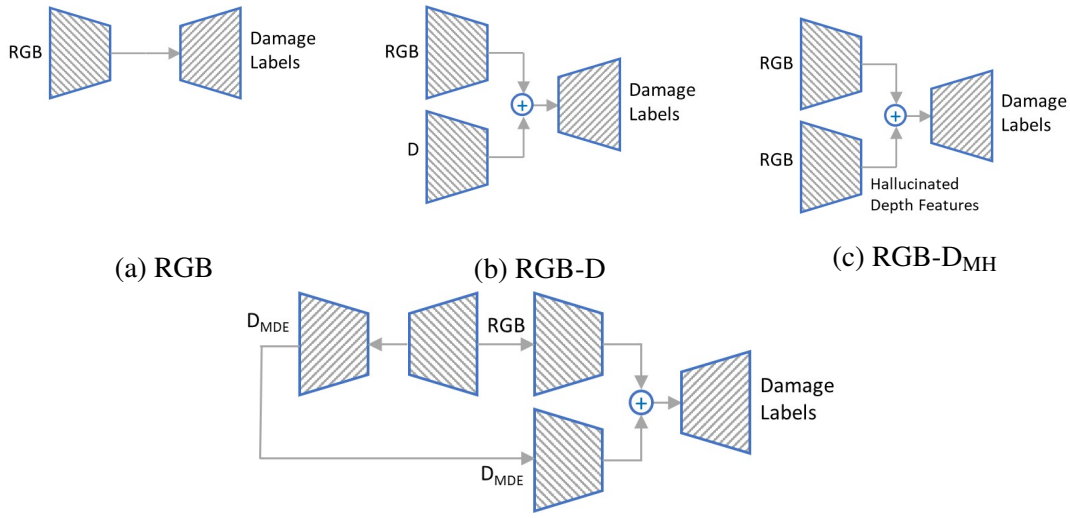
segmentation. This study uses absolute depth and surface normal maps to represent depth data. The fusion of depth data was observed to enhance the segmentation performance significantly. Additionally, two surrogate techniques are proposed to avoid depth sensing at test time and yet retain the benefits of depth fusion, as shown in Figure 1. A pure RGB-based model is used as a reference for traditional convolutional neural network approaches (Figure 1(a)). This study also explores a fusion-based architecture (RGB-D) where a pair of encoders take RGB and encoded depth (D) as input (Figure 1(b)). The last decoder layers' feature maps are fused and sent to a shared decoder to obtain predicted damage labels. This fusion approach can be leveraged when depth sensing is enabled during testing. Additionally, a modality hallucination-based fusion scheme (RGB-D_{MH}) is explored (Figure 1(c)), which enables the simulation of mid-level convolutional D features from a single-frame RGB image. These hallucinated D features are fused with RGB features before being sent to a common decoder. Furthermore, the study examines a fusion strategy (RGB-D_{MDE}) where deep learning techniques simulate the encoded depth (D_{MDE}) data from corresponding RGB frames. D_{MDE} is then fused with RGB data in the same way as in the case of RGB-D. Altogether, the surrogate strategies (RGB-D_{MH} and RGB-D_{MDE}) require depth data only for model training. The need for depth sensing during testing is eliminated without significantly reducing segmentation performance. The proposed depth fusion framework is validated on a computer-generated synthetic dataset containing three damage categories commonly observed in reinforced concrete buildings subjected to seismic excitations: spalling, exposed rebars, and severely buckled rebars. Overall, this study makes several key contributions, including demonstrating that deep learning-based damage segmentation algorithms can significantly improve accuracy through the fusion of RGB and depth information. The study explores two different strategies for encoding depth data and proposes surrogate techniques that provide the benefits of depth fusion without requiring depth sensing at test time.

SYNTHETIC DATA GENERATION

The single biggest factor that deterred the scientific community from exploring the utility of depth data with regard to vision-based autonomous condition assessment of civil infrastructure is the scarcity of a publicly available damage dataset that contains depth information. This shortcoming is overcome in this study by using a rasterization-based game engine called Houdini to generate a database of synthetic damage data containing color and depth information. The database contained a total of 1792 scenes belonging to three different damage categories, namely, spalling, spalling with exposed rebars, and severely buckled rebars (Figure 2). The generated data was labeled automatically using a special feature inbuilt into Houdini.

DEPTH ENCODING TECHNIQUES

Representing the depth information in a proper way is paramount for getting the most out of depth fusion. The quest for a suitable strategy for representing depth data has led to the emergence of various encoding techniques such as absolute depth-based encoding (ADE) and surface normal-based encoding (SNE), which are considered in



(a) RGB (b) RGB-D (c) RGB-D_{MH}

(d) RGB-D_{MDE}

Figure 1. Various depth fusion strategies are explored in this study. The trapezoids tapered on the right and left represent encoders and decoders, respectively. The '+' sign symbolizes a fusion of convolutional features.

this study. In the ADE, the value stored at each pixel of the depth map (Figure 3(b)) represents the absolute distance between the camera and a physical point in the 3D. On the other hand, in the SNE (Figure 3(c)), the depth data is represented in terms of X, Y, and Z components of the surface normal vector computed at each point in the scene. The resulting surface normal map looks like a texture and provides valuable information about the presence of damage in the scene.

METHODOLOGY

This study utilized a baseline model involving a fully-convolutional encoder-decoder network (Figure 4). The encoder is based on the VGG-16 architecture [3] and extracts important features from the input image. On the other hand, the decoder upsamples those features to match the original input resolution, ensuring that the output segmentation mask corresponds pixel-to-pixel with the input image. The effectiveness of the proposed surrogate techniques is compared against a pure RGB-based model (Figure 4(a)) and an RGB-D fusion network (Figure 4(b)) that can be used when depth data are available. The fusion network has two encoders dedicated to the RGB and D modalities (Figure 4(b)). The feature maps from the last layers of the two encoders are merged before being passed to the shared decoder layers.

The modality hallucination technique improves the accuracy of a test-time RGB-only network by using absolute depth or surface normal data, denoted as D in this study, as side information during training. This technique requires paired RGB and D images during training and introduces a third encoder called the hallucination branch that takes RGB images as input (Figure 5(a)). A regression-based hallucination loss allows for efficient information sharing between the D and hallucination branches, as shown in Eq. 1.

$$\mathcal{L}_{hallucination} = \left\| \psi_l^D - \psi_l^H \right\|_2^2 \tag{1}$$

(a)

(b)

(c)

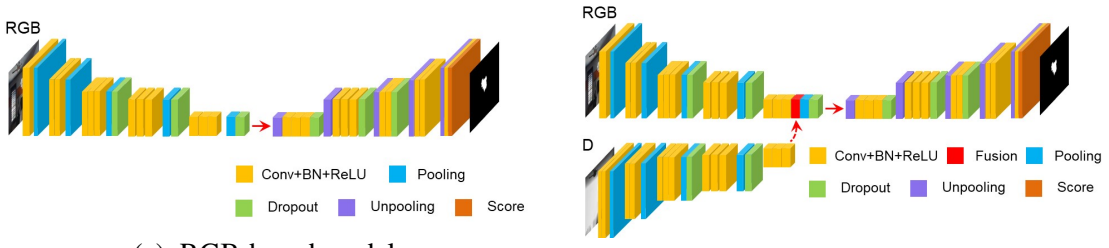
Figure 2. Damage categories considered in this study - (a) spalling, (b) spalling with exposed rebars, (c) spalling with buckled rebars.

(a) RGB image (b) Absolute depth map (c) Surface normal map

Figure 3. Various depth encoding techniques - (a) RGB image of the scene, (b) absolute depth-based encoding (ADE), (c) surface normal-based encoding (SNE).

where ψ_l^D and ψ_l^H are mid-level features from the D and hallucination branches, respectively. This loss is minimized alongside a standard supervised loss over the class labels to ensure that the mid-level convolutional features learned by the hallucination and D branches are similar. At the end of the training process, the D branch becomes redundant because the mid-level features generated by the D branch can now be generated by the hallucination branch using RGB data. During test time, the D branch can be discarded, and the mid-level activations from the hallucination branch can be fused to the RGB branch to create a more informed test-time RGB-based network (Figure 5(b)). This technique significantly outperforms a standard benchmark model trained solely on RGB data and eliminates the need for depth sensing without any loss of segmentation accuracy.

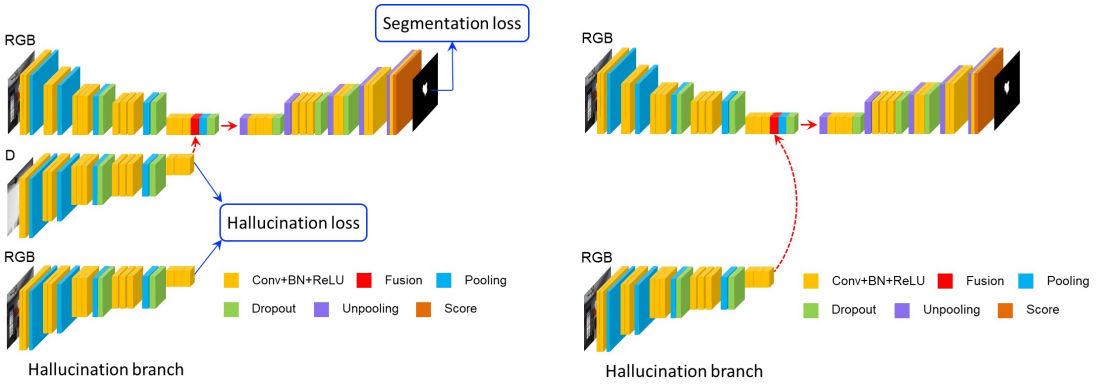
On the other hand, the main objective of monocular depth estimation is to predict depth values for each pixel of an RGB image. Recent advancements in deep learning techniques have shown encouraging results in predicting a dense depth map from a single RGB image. This study examined a convolutional neural network-based approach in this regard. The reconstructed depth maps are paired with the corresponding



(a) RGB-based model

(b) RGB-D fusion network

Figure 4. Network architectures that are used as benchmarks to evaluate the efficacy of the proposed surrogate techniques. D indicates absolute depth and surface normal maps for ADE and SNE, respectively.



(a) At training time

(b) At test time

Figure 5. The schema of modality hallucination. D indicates the absolute depth and surface normal maps for ADE and SNE, respectively.

RGB frames to be used as inputs for the fusion-based segmentation models in the case of ADE, whereas the SNE requires the depth images to be converted to surface normal maps before being fed to the fusion network. This study uses a standard encoder-decoder network with skip connections (Figure 6) to generate high-resolution depth maps from single frame RGB images. The encoder is taken from a DenseNet-169 architecture [4], which was pretrained on the ImageNet dataset [5]. The decoder consists of a series of up-sampling layers. The predicted depth values are compared to ground truth depths using a composite loss function that includes an L1 loss on the depth values, an L1 loss on the gradients of the depth image, and a structural similarity loss [6].

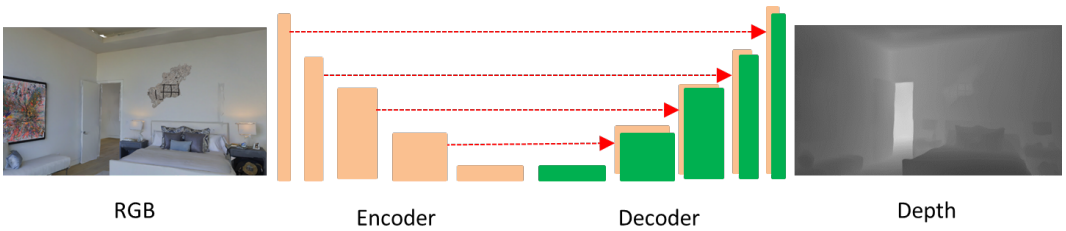


Figure 6. Deep learning-based monocular depth estimation.

RESULTS AND DISCUSSIONS

This study leveraged modality hallucination and monocular depth estimation techniques to create substitutes for actual depth sensing during testing. To evaluate their effectiveness, the intersection over union (IoU) between the predicted and target damage regions was computed. A five-fold cross-validation was conducted, and the mean class-specific IoUs were combined to create an overall IoU, as shown in TABLE I). The results showed that RGB-D_{MH} and RGB-D_{MDE} were comparable to RGB-D in terms of accuracy for ADE. However, for SNE, RGB-D_{MH} had a 4% lower IoU compared to RGB-D. Nevertheless, RGB-D_{MH} was still much better than a single-modality RGB-based model. Additionally, RGB-D_{MH} had higher segmentation accuracy than RGB-D_{MDE}, in case of SNE.

TABLE I. IoU mean values for different fusion architectures. D indicates absolute depth and surface normal maps for ADE and SNE, respectively.

	RGB	ADE			SNE		
	-	RGB-D	RGB-D_{MH}	RGB-D_{MDE}	RGB-D	RGB-D_{MH}	RGB-D_{MDE}
IoU Mean	0.690	0.880	0.874	0.873	0.932	0.891	0.876

On the other hand, the processing speed of RGB-D_{MH} was found to be faster than RGB-D and at par with pure RGB-based model (TABLE II). This is especially advantageous for SNE, where a lot of time is usually spent on surface normal estimation. On the other hand, the RGB-D_{MDE} technique takes a considerably longer time, particularly for SNE, due to its multi-stage processes. Overall, RGB-D_{MH} stands out as the best surrogate strategy in terms of accuracy and processing speed.

TABLE II. Processing time (seconds/image) for various fusion strategies. D indicates the absolute depth and surface normal maps for ADE and SNE, respectively.

RGB	ADE			SNE		
-	RGB-D	RGB-D_{MH}	RGB-D_{MDE}	RGB-D	RGB-D_{MH}	RGB-D_{MDE}
0.075	0.092	0.076	0.167	0.362	0.076	0.705

CONCLUDING REMARKS

This study shows that depth fusion can enhance the performance of a deep learning-based multi-class damage segmentation framework. A synthetic database is generated using computer graphics software containing three different damage categories that are commonly observed in reinforced concrete structures subject to extreme loading. Various encoding techniques are considered to represent depth data. Several experiments are conducted which suggest that the proposed fusion-based framework outperforms the traditional RGB-based approaches. The study also demonstrated that depth fusion can be achieved without requiring any physical depth-sensing at test time. To this end, two surrogate techniques based on modality hallucination and monocular depth estimation are explored to simulate depth information from the corresponding RGB frames. Results showed that modality hallucination is more accurate and considerably faster than

the monocular depth estimation-based approach. Not just that, its computational cost is comparable to a single modality RGB-based network and lower than a fusion model leveraging real depth measurements. Overall, this research paves the way for more resilient civil infrastructure systems through multimodal inspection. The scope for future work includes validating the proposed approach with real RGB-D data from various structural systems.

ACKNOWLEDGMENT

This study was supported in part by a fund from Bentley Systems, Inc.

REFERENCES

1. Ghosh Mondal, T., M. R. Jahanshahi, R.-T. Wu, and Z. Y. Wu. 2020. "Deep learning-based multi-class damage detection for autonomous post-disaster reconnaissance," *Structural Control and Health Monitoring*, 27(4):e2507.
2. Ghosh Mondal, T., M. R. Jahanshahi, and Z. Y. Wu. 2023. "Deep Learning-Based RGB-D Fusion for Multimodal Condition Assessment of Civil Infrastructure," *Journal of Computing in Civil Engineering*, 37(4):04023017.
3. Simonyan, K. and A. Zisserman. 2014. "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*.
4. Huang, G., Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. 2017. "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.
5. Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, Ieee, pp. 248–255.
6. Alhashim, I. and P. Wonka. 2018. "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*.