Earthq Eng & Eng Vib (2023) 22: 55-68 **DOI**: https://doi.org/10.1007/s11803-023-2155-2

Fusion of color and hallucinated depth features for enhanced multimodal deep learning-based damage segmentation

Tarutal Ghosh Mondal[1†] and Mohammad Reza Jahanshahi[2,3‡]

1. *Civil, Architectural and Environmental Engineering, Missouri University of Science and Technology, Rolla, MO, USA*
2. *Lyles School of Civil Engineering, Purdue University, West Lafayette, IN, USA*
3. *School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA*

Abstract: Recent advances in computer vision and deep learning have shown that the fusion of depth information can significantly enhance the performance of RGB-based damage detection and segmentation models. However, alongside the advantages, depth-sensing also presents many practical challenges. For instance, the depth sensors impose an additional payload burden on the robotic inspection platforms limiting the operation time and increasing the inspection cost. Additionally, some lidar-based depth sensors have poor outdoor performance due to sunlight contamination during the daytime. In this context, this study investigates the feasibility of abolishing depth-sensing at test time without compromising the segmentation performance. An autonomous damage segmentation framework is developed, based on recent advancements in vision-based multi-modal sensing such as modality hallucination (MH) and monocular depth estimation (MDE), which require depth data only during the model training. At the time of deployment, depth data becomes expendable as it can be simulated from the corresponding RGB frames. This makes it possible to reap the benefits of depth fusion without any depth perception per se. This study explored two different depth encoding techniques and three different fusion strategies in addition to a baseline RGB-based model. The proposed approach is validated on computer-generated RGB-D data of reinforced concrete buildings subjected to seismic damage. It was observed that the surrogate techniques can increase the segmentation IoU by up to 20.1% with a negligible increase in the computation cost. Overall, this study is believed to make a positive contribution to enhancing the resilience of critical civil infrastructure.

Keywords: multimodal data fusion; depth sensing; vision-based inspection; UAV-assisted inspection; damage segmentation; post-disaster reconnaissance; modality hallucination; monocular depth estimation

1 Introduction

1.1 Motivation

Recent research trends in structural health monitoring indicate that autonomous robotic platforms equipped with multimodal sensors and empowered by deep learning-based onboard processing capability will accompany or even replace human inspectors to automate the future inspection processes (Mondal *et al*., 2020; Mondal and Jahanshahi, 2020, 2022; Yeum *et al*., 2019). However, despite considerable research efforts and technological advances, the adoption of these automation-driven inspection solutions has not kept pace, mainly due to reliability issues. This pushback from the end-users has prompted the scientific community to enhance the accuracy and reliability of deep learning-based decision-making systems (Gao and Mosalam, 2022), which led to enhanced algorithmic complexity and computational cost. Meanwhile, state-of-the-art research in other disciplines brought to light that the accuracy of deep learning-based models can also be increased by enriching the information content of the input data. In this regard, a number of studies have looked into the fusion of RGB and depth information, which outperformed the conventional RGB-based deep learning approaches. Mondal (2021) indicated that depth data provides valuable structural information which complements the color information provided by RGB data, leading to improved segmentation accuracy. Schwarz *et al*. (2018) demonstrated that the efficiency of CNN-based robotic scene understanding and manipulation can be enhanced by infusion of depth information. Hazirbas *et al*. (2016) proposed a fusion-based CNN architecture to show that the incorporation of depth fusion can substantially improve the segmentation accuracy of indoor scenes. Park *et al*. (2017) proposed a multi-level feature fusion

Correspondence to: Mohammad Reza Jahanshahi, Lyles School of Civil Engineering, School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA
Tel: +1-765-494-2217; Fax: +1-765-494-0395
E-mail: jahansha@purdue.edu

†Post-Doctoral Fellow; ‡Associate Professor

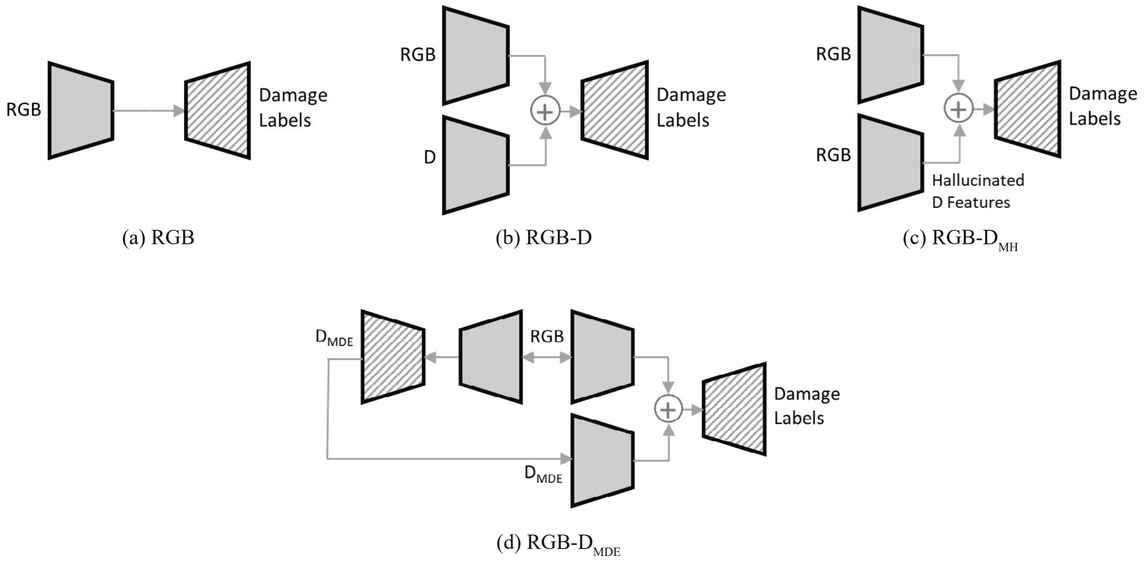
Received September 7, 2022; **Accepted** November 24, 2022

network to illustrate that a fusion of depth features increases the accuracy of indoor semantic segmentation. Cheng *et al*. (2017) leveraged a gated fusion of RGB-D features for enhanced semantic segmentation of indoor scenes. Xu *et al*. (2017) adopted a shared weights strategy and parameter-free correlation of modality-correlated and modality-specific features for RGB-D object detection, leading to an overall improvement in the detection accuracy. Ophoff *et al*. (2018) invoked a single-pass CNN architecture to fuse depth and visual sensor data for real-time pedestrian detection resulting in an improved accuracy. Notwithstanding, the structural health monitoring community has been a laggard on exploring this important research area. Lately, Alexander *et al*. (2022) fused RGB and thermal images for enhanced deep learning-based crack detection in civil infrastructure. Besides, Zhou and Song (2020) investigated the fusion of intensity and range images for CNN-based classification of roadway cracks. Wang *et al*. (2022) developed a synthetic robotic system capable of automated visual surveillance of construction sites leveraging RGB-D fusion. A few other studies, on the other hand, resorted to RGB-D fusion for the transformation of crack width estimated by vision-based techniques from pixels to actual physical units (Kim, 2021). However, many knowledge gaps still exist, which call for increased attention from the scientific community in coming times. This study aims to address one such important knowledge gaps by focusing on the fusion of RGB and depth information enabling deep learning-based enhanced multimodal defect segmentation in reinforced concrete buildings.

Despite proven advantages of depth fusion, it is not to be forgotten that depth sensors are not yet as pervasive and ubiquitous as RGB cameras. Moreover, practical application of depth sensing during real robotic inspection has many challenges. The traditional lidar-based depth sensors are generally large and weighty, and therefore not suitable to be integrated with mobile robotic platforms. The recent consumer-grade depth sensors, on the other hand, have the advantages of being lightweight and low-cost. However, many of these sensors exploit laser scanning techniques which are susceptible to interference by sunlight, leading to a poor outdoor performance. Besides, depth sensors may lead to an increased energy consumption reducing the operating life of unmanned aerial vehicles (UAVs), which rely on on-board batteries as primary energy sources. This reduces the efficiency of robotic inspection by increasing the inspection time and costs.

1.2 Contributions

In view of these practical constraints, it is regarded ideal to forego depth sensing at test time without foregoing the benefits of depth fusion. This study aims to achieve this research objective by leveraging two important advances in the area of multi-modal sensing, namely MH (Hoffman *et al*., 2016; Gunasekar *et al*., 2020) and MDE (Bhoi, 2019; Zhao *et al*., 2020). A fully-convolutional encoder-decoder network is used as a baseline model to assess the performance of the surrogate techniques. The depth data is represented in this study in the form of absolute depth or surface



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

(d) RGB-D_{MDE}

Fig. 1 Various depth fusion strategies are explored in this study. The encoded depth data is denoted by D. The trapezoids tapered on the right and left represent encoders and decoders, respectively. The '+' sign symbolizes a fusion of convolutional features. RGB implies a pure RGB-based model. RGB-D indicates a feature level fusion of RGB and measured D data. RGB-D_{MH} signifies the fusion of RGB features and D features hallucinated from RGB image. Finally, RGB-D_{MDE} connotes the feature level fusion of RGB and D data estimated from the corresponding RGB frame through monocular depth estimation

normal maps (Section 2.2). Additionally, four different fusion strategies are investigated, as shown in Fig. 1. A pure RGB-based model is used as a representative of traditional CNN approaches (Fig. 1(a)). Next, a fusion-based architecture (RGB-D) is invoked comprising a pair of encoders taking RGB and encoded depth (D) as input (Fig. 1(b)). The feature maps produced by the last decoder layers are fused and sent to a shared decoder to obtain the predicted damage labels. This fusion approach can be leveraged when depth sensing is enabled at test time. In addition to this, an MH-based fusion scheme (RGB-D_{MH}) is explored, which enables simulation of mid-level convolutional D features from a single-frame RGB image (Fig. 1(c)). These hallucinated D features are fused with RGB features before being sent to a common decoder. More details about this fusion scheme are included in Section 3.1. Further, this study looked into a fusion strategy (RGB-D_{MDE}) where the encoded depth (D_{MDE}) data are simulated from the corresponding RGB frames exploiting deep learning techniques. D_{MDE} is then fused with RGB data exactly in the same manner as in the case of RGB-D. Section 3.2 provides additional details about this fusion approach. It should be noted in this context that this study considered only feature-level fusion as it is demonstrated by previous studies (Mondal, 2021) to be superior to image-level fusion. Altogether, the surrogate strategies (RGB-D_{MH} and RGB-D_{MDE}) lead to a situation where depth data are required only for model training. The need for depth sensing at test time is eliminated without considerably undermining the 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, namely spalling, exposed rebars, and severely buckled rebars. The textural similarities between different damage categories made the traditional approach of relying solely on RGB data less rewarding, making the fusion of depth information all the more critical. Overall, the key contributions of this study can be summarized as follows:

- This study demonstrates that the accuracy of deep learning-based damage segmentation algorithms can be significantly improved by the fusion of RGB and depth information.
- Two different depth encoding techniques are explored for the representation of the depth data.
- It is shown that depth sensing is not indispensable at test time. A pair of surrogate techniques are investigated, which eliminate the need for depth sensing at test time without foregoing the benefits of depth fusion.
- The proposed framework is validated on computer-generated RGB-D data containing three different damage categories commonly encountered in reinforced concrete buildings subjected to extreme loading.

2 Data preparation

2.1 Data generation using computer graphics techniques

Unfortunately, there is no publicly available damage dataset that can provide depth data for scientific investigations on RGB-D fusion. As a workaround, this study leveraged a 3D animation and visual effects software called Houdini (Elkins, 2020) for the generation of photo-realistic RGB-D data. 3D reconstructions of real buildings provided by Matterport3D dataset (Chang *et al*., 2017) are used as baseline models (Fig. 2(a)). A region of interest is manually demarcated and isolated from the remaining model for damage incorporation (Fig. 2(d)). The damage is induced by Boolean subtraction of a solid object (Fig. 2(e)) from the isolated wall section (Figs. 2(g)-2(h)). The solid object can be created simply by noise addition to a rudimentary geometric entity such as a sphere or an ellipsoid (Fig. 2(b)). The isolated and damage part of the wall is then merged with the remaining structure to retrieve the entire building model (Fig. 2(k)). At the same time, a mesh of rebars is created and placed on the damaged part of the model (Fig. 2(f)). Finally, the steel rebars and the damaged concrete surface are textured (Figs. 2(c) and 2(j)) to produce an appearance of concrete spalling with exposed rebars (Fig. 2(l)). In some cases, the mesh of rebars is suitably deformed to mimic buckled reinforcement bars. Various modeling parameters such as rebar diameter and spacing are ensured to be consistent with the guidelines of ACI 318 (Standard, 2011). The data generation process was guided by the observation of real damages rendered by several past earthquakes, such as Nepal earthquake, Ecuador earthquake, etc. (Shah *et al*., 2015; NCREE, 2016; Sim *et al*., 2016). The generated dataset contained wide variations in terms of shape, size, location, and texture of damage. However, one of the limitations of this data generation technique is that it requires significant human involvement. Future studies should endeavor to integrate the entire procedure into an open-source toolbox to increase the practicability and academic impact. Altogether, the generated dataset comprised 629 of such 3D scenes representing three different damage categories such as spalling, exposed rebars, and severely buckled rebars (Fig. 3). It should be noted here that this study did not consider concrete cracks, even though it is the most common damage type observed in reinforced concrete structures. This is because the thickness of concrete cracks, in many cases, is too small to be captured by inexpensive consumer-grade depth cameras, which are typically constrained by limited spatial resolution. Moreover, crack-induced depressions on concrete surfaces are generally not very significant and, more often than not, get overshadowed by the measurement noise present in depth data. As a

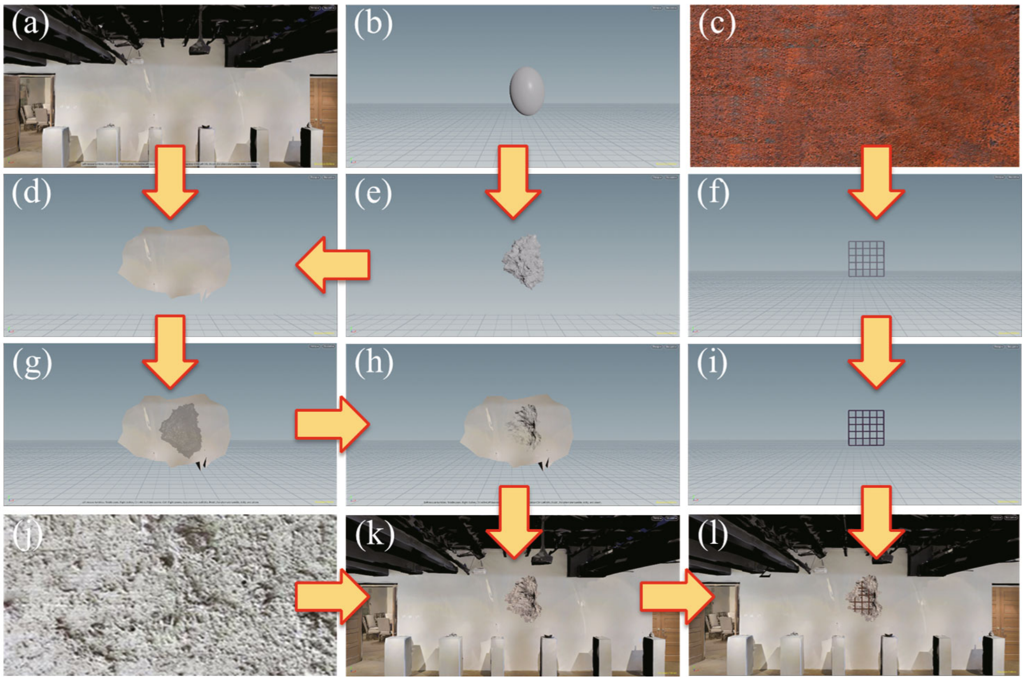
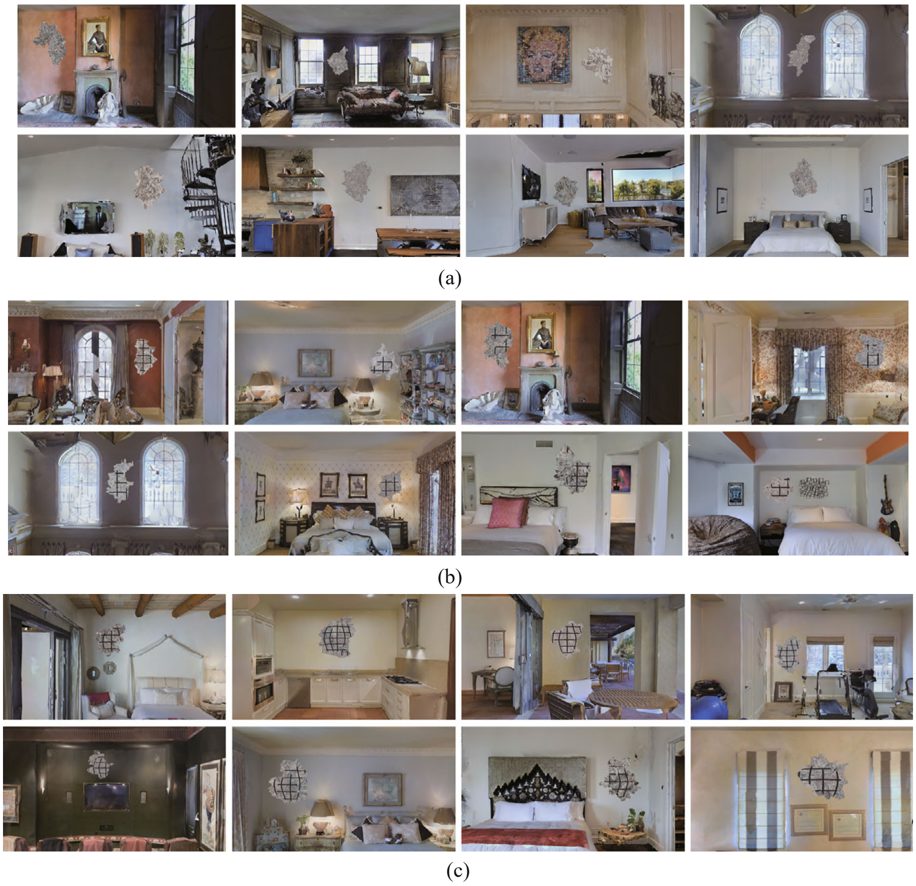


Fig. 2 3D photo-realistic damage data generation pipeline using computer graphics technique: (a) 3D reconstruction of real building, (b) solid ellipsoid, (c) texture map used to represent steel reinforcement bars, (d) isolated region of interest where damage is to be induced, (e) solid ellipsoid deformed due to noise induction, (f) mesh of rebars, (g) intersection of the isolated wall surface by the deformed ellipsoid, (h) material removal from the intersection zone, (i) texture applied to the rebar mesh, (j) texture map used to represent damaged concrete, (k) merger of the isolated wall section with the rest of structure, and (l) application of the texture shown in (j) on the damaged part and placement of the rebar mesh, resulting in a photo-realistic damaged building model


(a)

(b)

(c)

Fig. 3 Damage categories considered in this study: (a) spalling, (b) spalling with exposed rebars, (c) spalling with buckled rebars

result, concrete cracks usually behave as 2D structures where depth sensing fails to provide any meaningful information. Preliminary investigations by the authors also validate this hypothesis. Each scene is rendered from multiple camera positions and orientations resulting in 1,789 sets of paired RGB-D images. To mimic real world depth sensing, the depth data generated by the computer graphics software are contaminated by adding a small amount of Gaussian noise, which is consistent with the noise level typically exhibited by the first-generation Microsoft Kinect sensor (Zennaro *et al*. 2015). The dataset was automatically labelled by the computer graphics tool, which saved considerable time and human effort which are otherwise necessitated by manual annotation process. To test the generalization ability of the trained models, five-fold cross-validation was conducted in this study. At each cross-validation round, 90% of the generated data were used for training, and the remaining 10% data were used for testing (Fig. 4). It was ensured that the training and test data sets at a given cross-validation round did not share different views of the same damage scenario.

2.2 Depth data encoding

Previous studies have indicated that to get the best out of depth fusion, it is important to represent depth data in a proper way. In light of this, this study explored two different depth encoding techniques, namely absolute depth-based encoding (ADE) and surface normal-based encoding (SNE). In ADE, the depth at a point is represented by the absolute distance between the camera and the physical point projected on the principal axis of the camera. Figure 5(b) shows a typical depth map where the brighter pixels are relatively farther from the camera and the darker pixels are closer to the camera. On the other hand, in SNE, the depth at a point is represented by the surface normal vector at that point. Given the focal length of the camera and the depth map of a scene, the 3D position of each point in the scene can be computed making use of a pinhole camera model. Following this, the surface normal vector at each point can be estimated by fitting a local plane at respective 3D points with the help of other points in their immediate neighborhoods. The three components of the unit normal vector at each point are used to encode the depth information at respective locations. This leads to a surface normal map (Fig. 5(c)) comprising three channels corresponding to three components of the surface normal vectors. As the figure depicts, the points lying on a plane have a common surface normal vector and are therefore represented by the same color. In the damage region, however, the surface normal vectors are all over the place, and the uniformity in the texture is lost. This becomes a telltale sign of the presence of a damage in the scene.

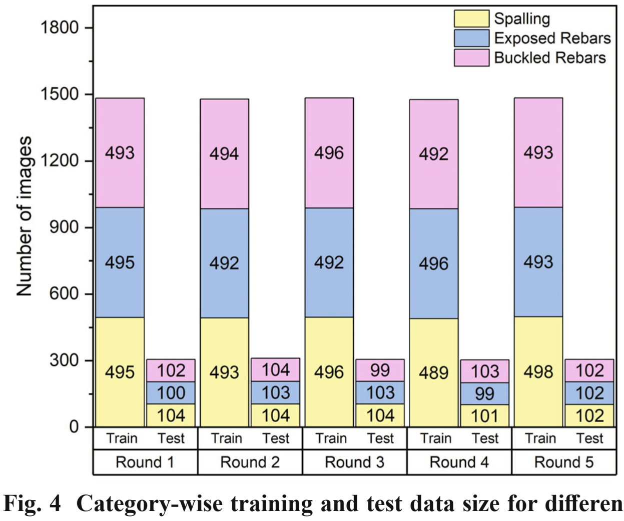


Fig. 4 Category-wise training and test data size for different cross-validation rounds

3 Methodology

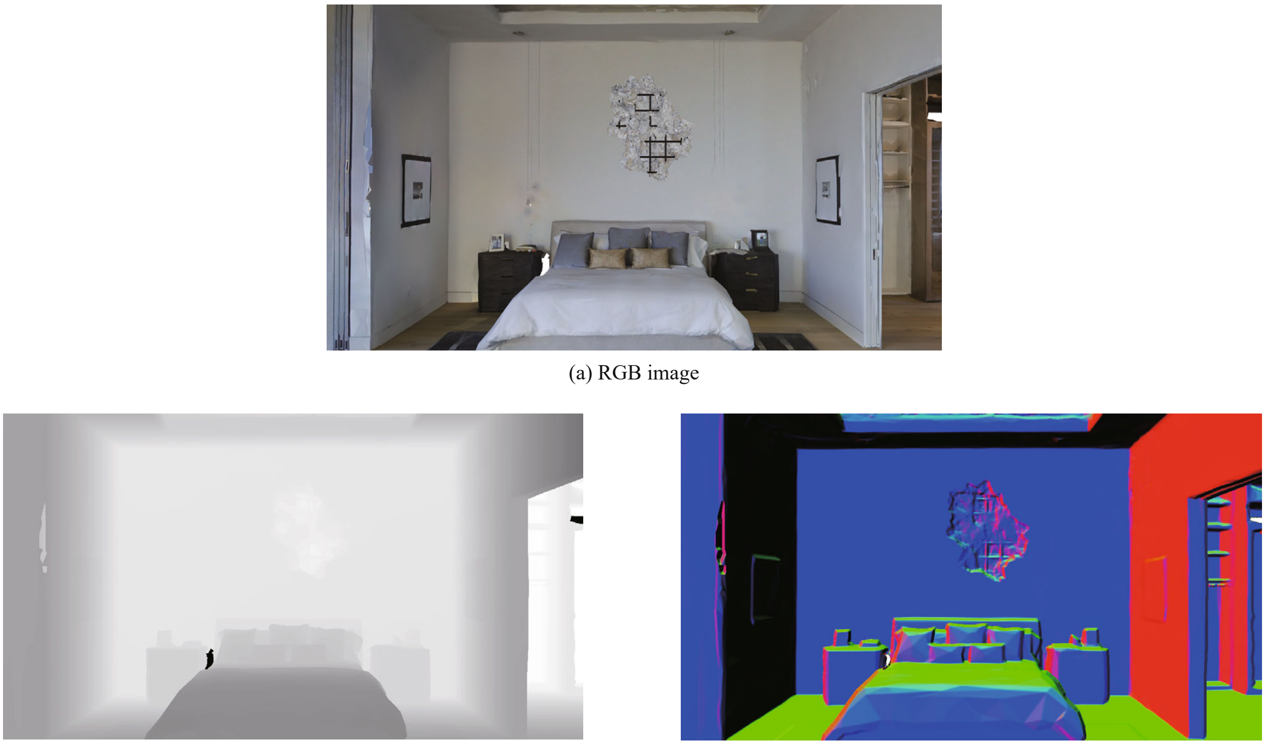
A fully-convolutional encoder-decoder network is used in this study as the baseline model (Fig. 6). The encoder extracts informative features from the input image and is adopted from the VGG-16 architecture (Simonyan and Zisserman, 2014). The decoder upsamples the features to the original input resolution. This ensures that the output segmentation mask has pixel-to-pixel correspondence with the input image. The surrogate techniques, which are introduced in the following sections, can be exploited when depth sensing is not viable at test time. The efficiency of these techniques are benchmarked against a pure RGB-based model (Fig. 6(a)) and an RGB-D fusion network (Fig. 6(b)) that can be used when depth data are available at test time. This fusion network has two encoders dedicated to the RGB and D modalities. The feature maps from the last layers of the two encoders are fused before being passed on to the shared decoder layers.

3.1 Modality hallucination (MH)

MH is a surrogate technique that uses absolute depth or surface normal data (denoted by D in this study) at training time as side information to produce a more informed test-time RGB only network. In this technique, an access to paired RGB and D images is presumed at the training time. Apart from the usual RGB and D branches, a third encoder, known as the hallucination branch, is also introduced (Fig. 7(a)), which takes RGB images as input. A regression-based hallucination loss is introduced to facilitate an efficient information sharing between the two modalities, as shown in Eq. (1).

$$L_{\text{hallucination}} = \| \psi^{\mathrm{D}} - \psi^{\mathrm{H}} \|_2^2 \qquad (1)$$

where ψ^{D} and ψ^{H} are mid-level features from the D and hallucination branches, respectively. This loss is

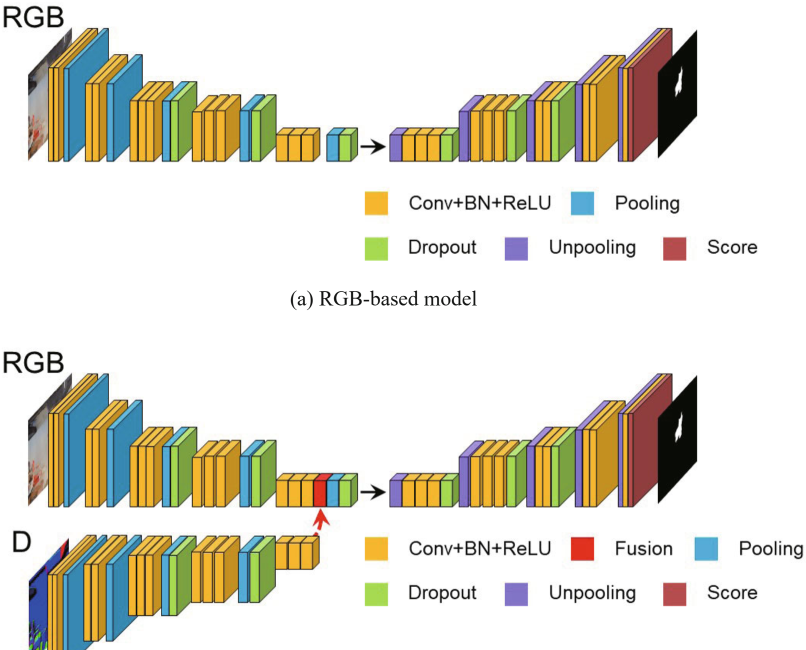


(a) RGB image

(b) Absolute depth map

(c) Surface normal map

Fig. 5 Various depth encoding techniques: (a) RGB image of the scene, (b) absolute depth-based encoding (ADE), (c) surface normal-based encoding (SNE)



(a) RGB-based model

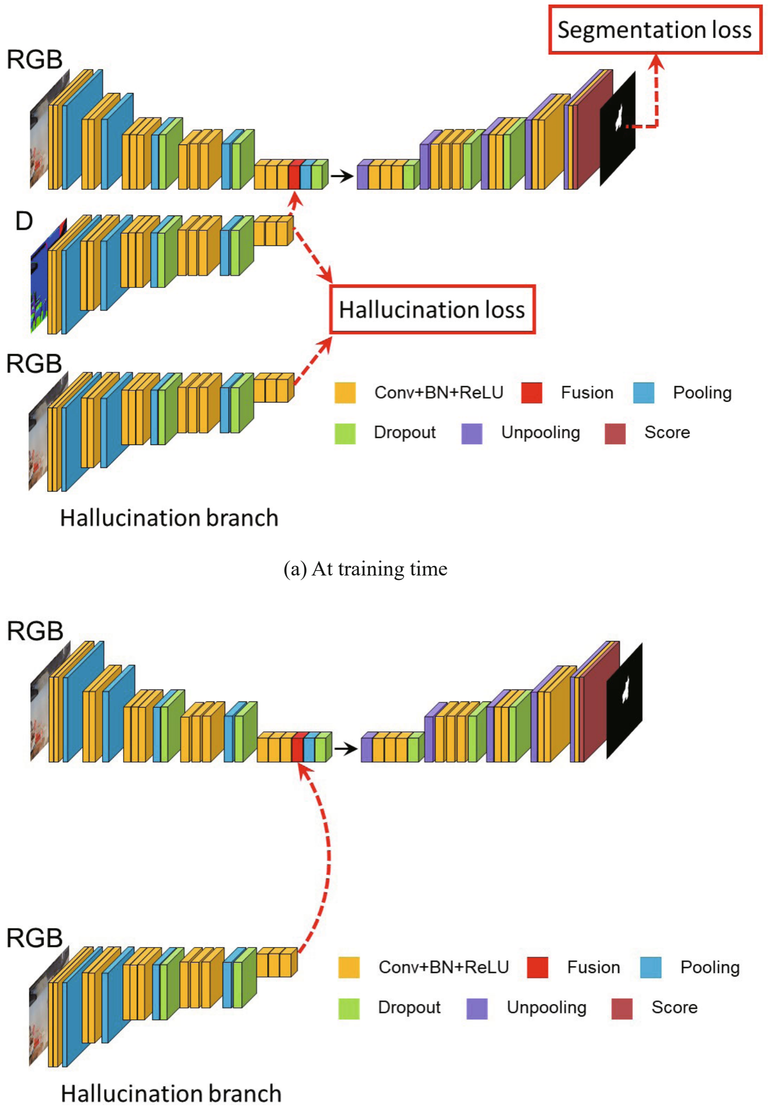
(b) RGB-D fusion network

Fig. 6 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

minimized alongside a standard supervised loss over the class labels, ensuring that the mid-level convolutional features learned by the hallucination and D branches mirror each other. In consequence, the D branch becomes redundant at the end of the training process. This is because the same mid-level features, which were hitherto generated by the D branch, can now be hallucinated by the hallucination branch using RGB data as network input. Thus, at 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 emulate a multi-modal fusion (Fig. 7(b)). This gives rise to a more informed test-time RGB-based network which significantly outperforms a standard benchmark model trained solely on RGB data, as illustrated in Section 4. This eliminates the need for depth sensing at test time without any appreciable loss of segmentation accuracy.

3.2 Monocular depth estimation (MDE)

The goal of MDE is to predict pixel-wise depth values corresponding to a given RGB image. Traditional depth estimation methods such as structure from motion (Özyeşil *et al*., 2017; Ullman, 1979; Wu *et al*., 2011; Schonberger and Frahm, 2016) and stereo matching (Cao *et al*., 2015; Zou and Li, 2010; Lazaros *et al*., 2008) rely on multiple views of a scene to generate a sparse depth map. However, many real-time inspection applications require depth map to be estimated from a single viewpoint. The recent developments in deep learning-based computer vision techniques have shown great promise of enabling this challenging task by predicting a dense depth map from a single frame RGB image in an end-to-end manner. This study explored two different approaches to this end based on convolutional neural network (CNN) and generative adversarial network (GAN). In the case of ADE, the reconstructed depth maps are paired with the corresponding RGB frames to be used as inputs for the fusion-based segmentation models. However, the SNE requires the depth images to be converted to surface normal maps before being fed to the fusion network.



(a) At training time

(b) At test time

Fig. 7 The schema of modality hallucination (MH). The network is trained to counterfeit intermediate D features from input RGB image, which makes depth sensing redundant at test time. D indicates absolute depth and surface normal maps for ADE and SNE, respectively

3.2.1 CNN-based approach

A standard encoder-decoder-based CNN with skip connections (Fig. 8) is used to predict detailed high-resolution depth maps from single frame RGB images (Table 1). The encoder is borrowed from the DenseNet-169 architecture (Huang *et al*., 2017) pre-trained on ImageNet dataset (Deng *et al*., 2009). The decoder, on the other hand, comprises a series of up-sampling layers. The baseline architecture is adopted from Alhashim and Wonka (2018) with some modifications. In the original study, the resolution of the final output depth maps was half the input resolution. However, the fusion strategies proposed in this study require that the input RGB and depth images have the same resolution. To address this specific need, this study appended an additional upsampling layer at the end of the network to ensure that the output resolution matches that of the input. The predicted depth values are regressed to ground truth depths by minimizing a composite loss function consisting of an L1 loss defined on the depth values, an L1 loss defined over the gradients of depth image, and a structural similarity loss (Alhashim and Wonka, 2018). The efficiency of this approach is discussed in Section 4.1.

3.2.2 GAN-based approach

A number of studies (Groenendijk *et al*., 2020; Kwak and Lee, 2020; Lore *et al*., 2018; Kumar *et al*., 2018; Tan *et al*., 2019), on the other hand, resorted to GAN for MDE. A GAN consists of a pair of neural networks known as the generator and the discriminator, which compete with each other. The generator is like a counterfeiter who tries to generate some fake depth images, and the discriminator is like the cop who tries to catch the counterfeiter. In the training phase, the generator becomes better and better at producing more realistic depth images until it can produce a perfect depth image, which fools the discriminator into believing that it is a real depth image. The same encoder-decoder network described in Section 3.2.1 is used in this study as a generator to produce artificial depth maps, which are then classified by a discriminator as real or fake (Fig. 9).

Table 1 Network architecture for the CNN used in MDE. The encoder is adopted from the DenseNet-169 (Huang *et al.*, 2017). The upsampling layers incorporate bilinear upsampling. Each Conv(B) convolutional layer is followed by a leaky ReLU activation

Layer	Output	Operation
Input	432×768×3	–
Conv(1)	216×384×64	DenseNet Conv1
Pool(1)	108×192×64	DenseNet Pool1
Pool(2)	54×96×128	DenseNet Pool2
Pool(3)	27×48×256	DenseNet Pool3
Conv(2)	13×24×1664	Convolution 1×1 of DenseNet Block4
Up(1)	27×48×1664	Upsample 2×2
Concat(1)	27×48×1920	Concatenate Pool3
Up(1)-Conv(A)	27×48×832	Convolution 3×3
Up(1)-Conv(B)	27×48×832	Convolution 3×3
Up(2)	54×96×832	Upsample 2×2
Concat(2)	54×96×960	Concatenate Pool2
Up(2)-Conv(A)	54×96×416	Convolution 3×3
Up(2)-Conv(B)	54×96×416	Convolution 3×3
Up(3)	108×192×416	Upsample 2×2
Concat(3)	108×192×480	Concatenate Pool1
Up(3)-Conv(A)	108×192×208	Convolution 3×3
Up(3)-Conv(B)	108×192×208	Convolution 3×3
Up(4)	216×384×208	Upsample 2×2
Concat(4)	216×384×272	Concatenate Conv1
Up(4)-Conv(A)	216×384×104	Convolution 3×3
Up(4)-Conv(B)	216×384×104	Convolution 3×3
Up(5)	432×768×104	Upsample 2×2
Conv3	432×768×1	Convolution 3×3

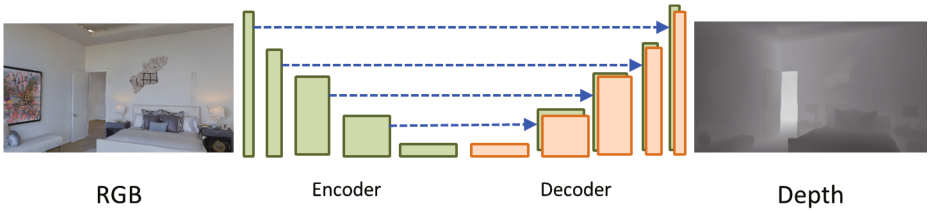


Fig. 8 CNN-based monocular depth estimation

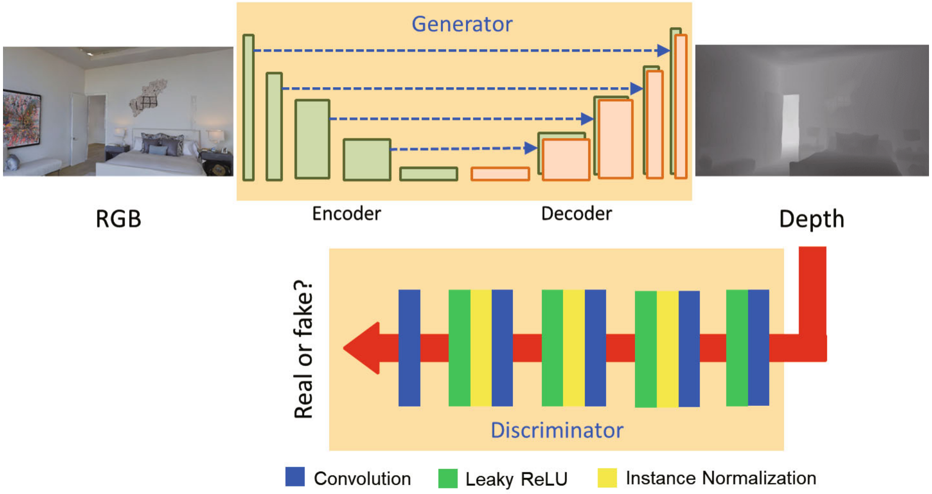


Fig. 9 GAN-based monocular depth estimation

The discriminator in this study, which facilitated this adversarial training, is adopted from the classical CycleGAN paper (Zhu *et al*., 2017). The performance of the GAN-based approach is described in the following section.

4 Results and discussions

This section discusses the results of the deep learning techniques presented in Section 3. First, the performance of the CNN and GAN-based MDE is evaluated. Subsequently, the efficiency of MH and MDE-based surrogate approaches is assessed, and the best strategy is identified based on accuracy and processing speed.

4.1 Comparing the Performance of CNN and GAN-based MDE

This section presents the results of CNN- and GAN-based MDE for depth estimation from a single RGB frame. Traditional deep learning algorithms require that the input data are suitably normalized before being fed into a deep learning model. Therefore, the ground truth and the estimated depth values were normalized between 0 and 1 in this study before computing the depth estimation accuracy. After this normalization, the estimated depth values were compared with the ground truth depths, and the average root mean square errors for five-fold cross-validations were observed to be 0.0435 and 0.0452 for the CNN- and GAN-based approaches, respectively. This indicates that adversarial training was not of any significant help, and therefore was not considered for any subsequent analysis. A few examples of the depth maps generated by the CNN- and GAN-based approaches are shown in Fig. 10 side by side with the corresponding RGB and ground truth depth images to demonstrate the efficiency of this technique.

4.2 Comparing the Performance of MH and MDE-based Surrogate Techniques

The main purpose of invoking MH and MDE was to create proxies for real depth sensing at test time. The efficiencies of these techniques were evaluated in terms of intersection over the union (IoU) between the predicted and target damage regions. The overall IoU, which is the mean of class-specific IoUs, are shown in Fig. 11. The small squares inside the rectangular boxes represent the mean overall IoU values produced by all cross-validation rounds. On the other hand, the horizontal lines inside the boxes represent the median values. The upper and lower sides of the rectangular boxes denote one standard deviation on either sides of the mean values. Last but not least, the whiskers protruding out of the boxes represent the minimum and maximum values. It is observed that, in the case of ADE, RGB-D$_{MH}$ and

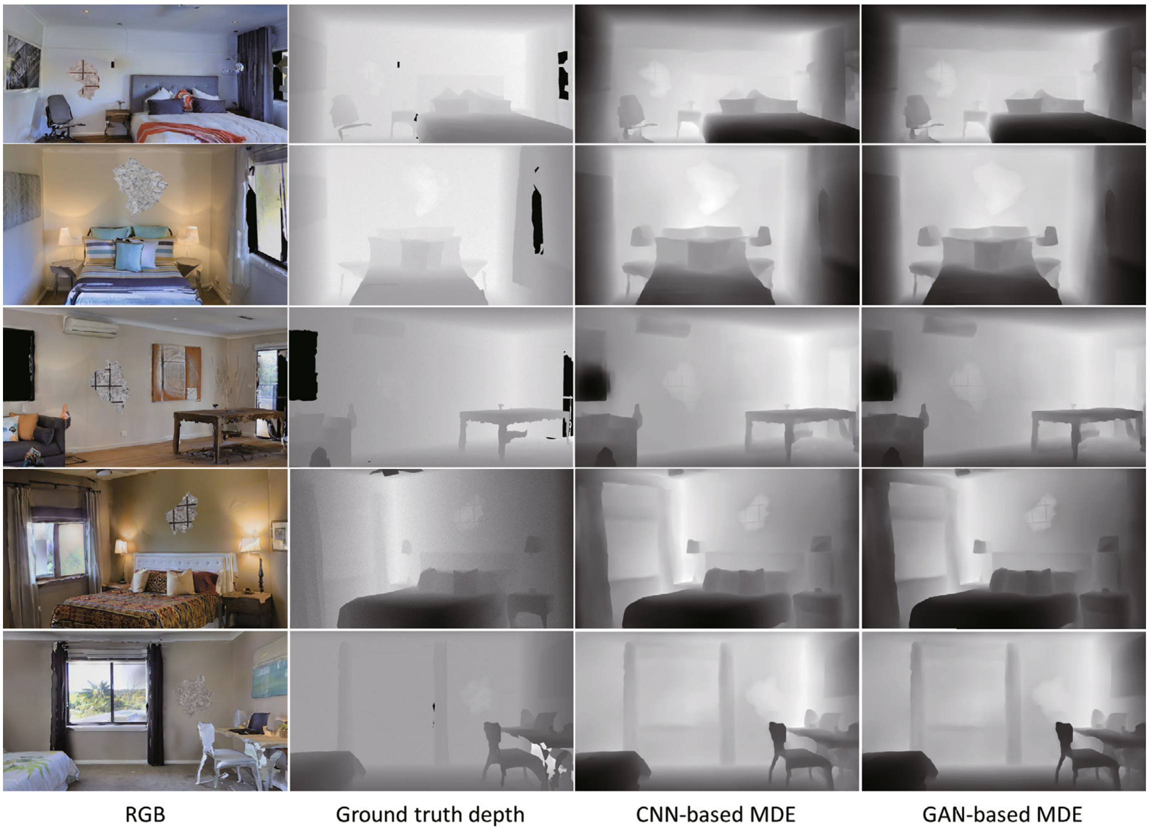


Fig. 10 Examples of monocular depth estimation (MDE) using CNN and GAN-based approaches

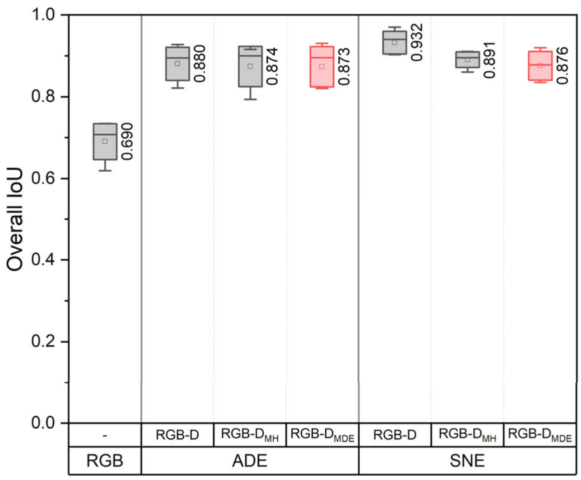


Fig. 11 Accuracy of RGB-D$_{\mathrm{MH}}$ and RGB-D$_{\mathrm{MDE}}$ as compared to RGB-D. D indicates absolute depth and surface normal maps for ADE and SNE, respectively

RGB-D$_{\mathrm{MDE}}$ have comparable accuracies, both being in the same ballpark with the RGB-D approach. In the case of SNE, however, RGB-D$_{\mathrm{MH}}$ suffered a 4% drop in IoU vis-à-vis the RGB-D approach. Nonetheless, this IoU is still streets ahead of that of a single-modality RGB-based model. Besides, RGB-D$_{\mathrm{MH}}$ demonstrated a clear edge over RGB-D$_{\mathrm{MDE}}$ for this encoding technique in terms of segmentation accuracy and robustness as measured by the standard deviation of IoU values (Table 2).

On the other hand, in terms of processing speed, it was observed (Fig. 12) that RGB-D$_{\mathrm{MH}}$ offers a major advantage for both ADE and SNE. It requires a processing time that is even lower than the RGB-D network and is at par with a pure RGB-based model on an NVIDIA Quadro RTX 8000 GPU. This can be attributed to the additional preconditioning (e.g., normalization of image tensor) of depth input that is necessitated by RGB-D but obviated by the RGB-D$_{\mathrm{MH}}$ approach. It is particularly advantageous for SNE, where

Table 2 IoU mean and standard deviation 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$_{\mathrm{MH}}$	RGB-D$_{\mathrm{MDE}}$	RGB-D	RGB-D$_{\mathrm{MH}}$	RGB-D$_{\mathrm{MDE}}$
IoU Mean	0.690	0.880	0.874	0.873	0.932	0.891	0.876
IoU Std. Dev.	0.044	0.041	0.050	0.049	0.027	0.019	0.035

considerable time (0.316 seconds/image on average) is expended in surface normal estimation from raw depth measurements. This step becomes inessential when MH is invoked. On the downside, it requires a training time that is 2.6 and 1.3 times higher than the RGB and RGB-D networks, respectively. Nevertheless, it leads to a win-win situation on many counts as it increases the accuracy at no additional cost of test-time processing speed. However, the RGB-D$_{\text{MDE}}$ technique requires considerably higher processing time, more so in the case of SNE. This can be attributed to the multi-stage process involving depth prediction, surface normal estimation, and semantic segmentation. Therefore, in the overall analysis, it can be concluded that RGB-D$_{\text{MH}}$ has a comparative advantage over RGB-D$_{\text{MDE}}$ in terms of both accuracy and processing speed, particularly in the case of SNE.

This study went a step further and looked into the class-specific IoU as depicted in Fig. 13. It was observed that RGB-D$_{\text{MH}}$ does not provide any significant advantage for the segmentation of spalling. However, when it comes to exposed or buckled rebars, RGB-D$_{\text{MH}}$ turns out to be a huge benefactor, irrespective of the encoding technique. It even exceeds the test-time performance of the RGB-D approach in case of these two damage categories by ably compensating for the lack of D data, as demonstrated qualitatively in Fig. 14. This can be attributed to the auxiliary unsupervised reconstruction loss, the minimization of which acts as a regularizer to enhance the generalization performance of the network (Le *et al*., 2018). Altogether, this implies that depth-sensing at test time is not indispensable. On the contrary, depth sensing can be surrogated at test time by employing state-of-the-art MH techniques. The authors believe that this is a significant addition to the existing knowledge base and will go a long way to enhance the efficiency of robotic inspection in the time to come.

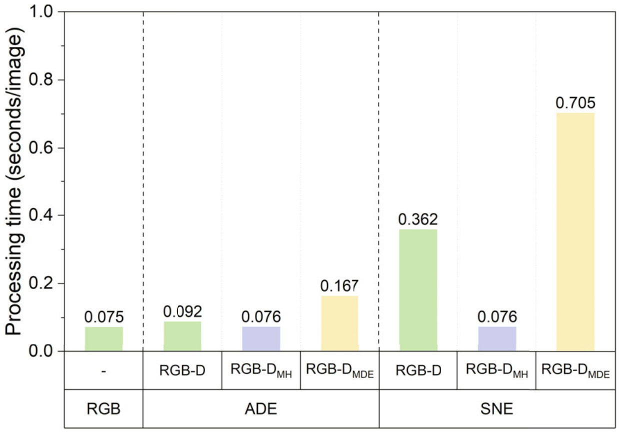
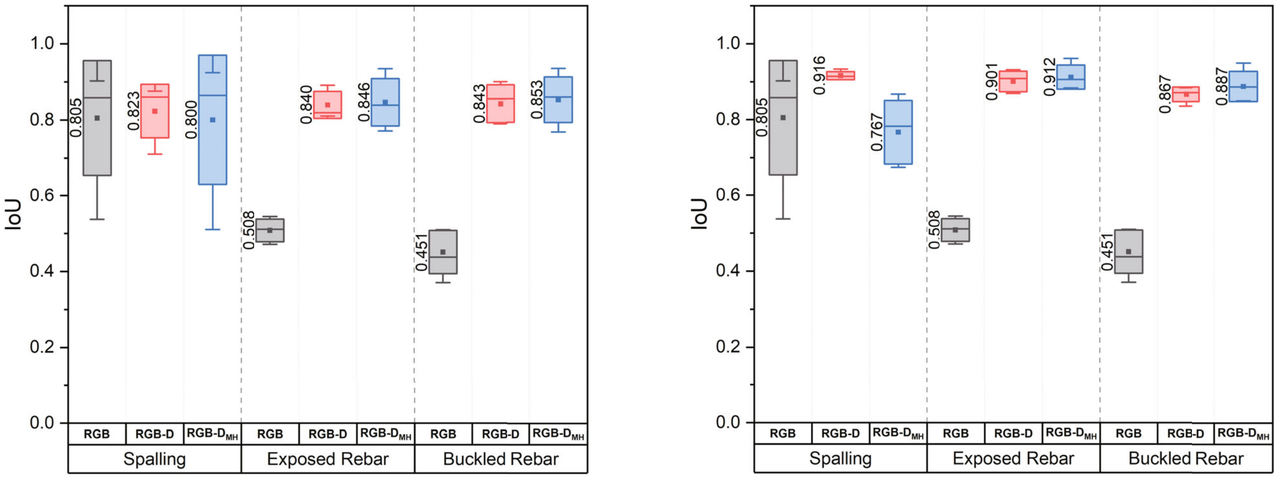


Fig. 12 Processing time for RGB-D$_{\text{MH}}$ and RGB-D$_{\text{MDE}}$ as compared to RGB-D. D indicates absolute depth and surface normal maps for ADE and SNE, respectively.

5 Conclusions

This study leveraged computer-generated visual inspection data to demonstrate that RGB-D fusion can be achieved without any test-time physical depth-sensing leading to a deep learning-based damage segmentation framework that is more accurate than the traditional RGB-based approaches. A couple of surrogate techniques based on MH and MDE are explored to concoct depth information at test time from the corresponding RGB frame. It was observed that RGB-D$_{\text{MH}}$ is more accurate than the RGB-D$_{\text{MDE}}$ approach. Not just that, it is even more accurate than the RGB-D approach in the case of relatively severe damage categories such as exposed and buckled rebars. In terms of processing speed also, it is faster than RGB-D$_{\text{MDE}}$ and even RGB-D, having a processing time comparable to a single-modality RGB-based network. On the whole, this study is hoped to



(a) ADE

(b) SNE

Fig. 13 Class-wise accuracy of RGB-D$_{\text{MH}}$ as compared to RGB-D and RGB-based networks. D indicates absolute depth and surface normal maps for ADE and SNE, respectively

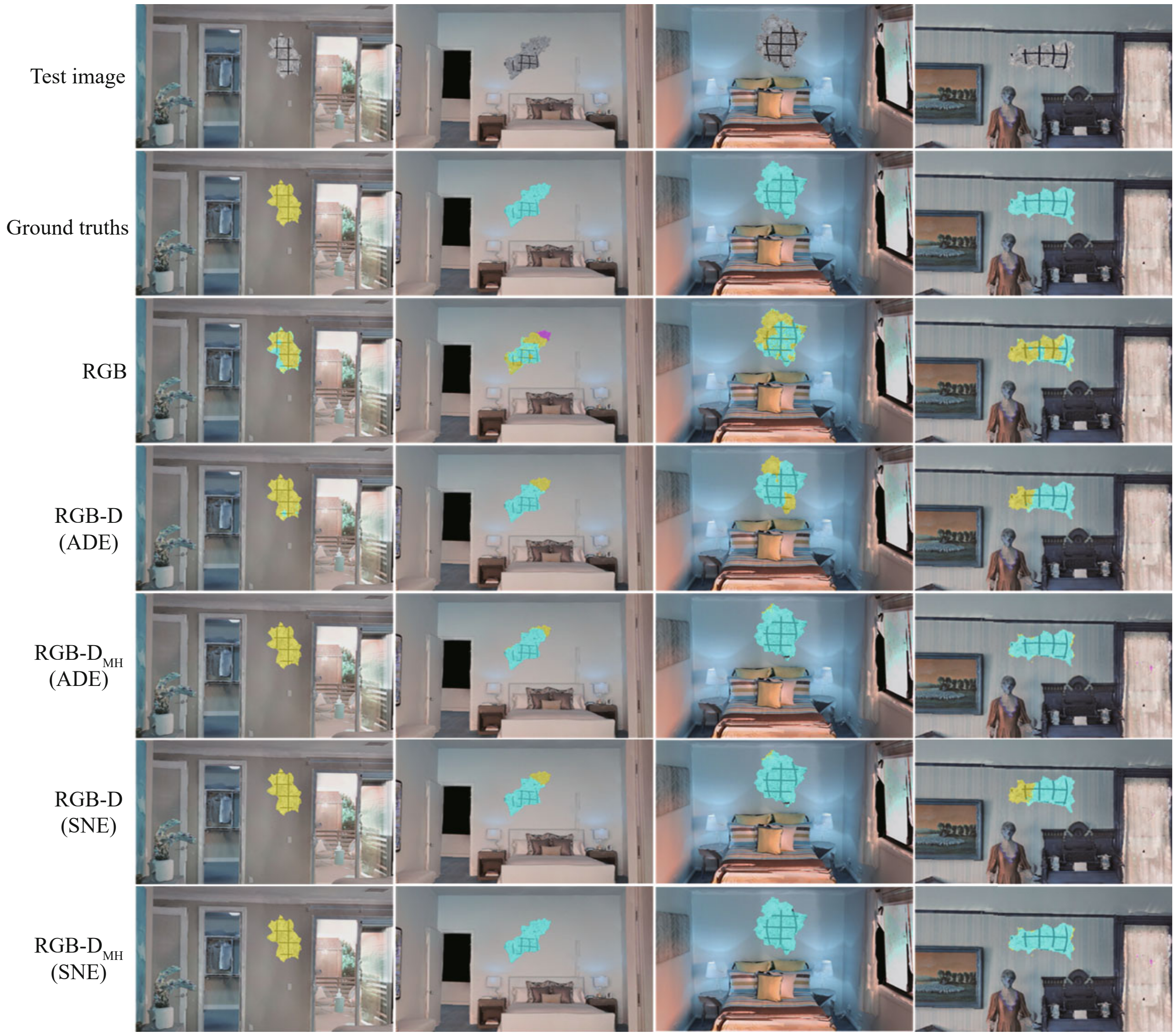


Fig. 14 Sample segmentation results. Magenta color denotes spalling, yellow color denotes exposed rebars, cyan color denotes buckled rebars. D indicates absolute depth and surface normal maps for ADE and SNE, respectively

blaze a new trail in multimodal inspection leading to more resilient civil infrastructure systems. Validation of the proposed approach with real RGB-D data from various structural systems is scope for future work.

Acknowledgement

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

References

ACI 318-11 (2011), *Building Code Requirements for Structural Concrete*, American Concrete Institute, USA.

Alexander QG, Hoskere V, Narazaki Y, Maxwell A, Spencer BF (2022), "Fusion of Thermal and RGB Images for Automated Deep Learning Based Crack Detection in Civil Infrastructure," *AI in Civil Engineering*, **1**(1): 1–10.

Alhashim I and Peter W (2018), "High Quality Monocular Depth Estimation via Transfer Learning," *arXiv preprint arXiv:1812.11941*.

Bhoi A (2019), "Monocular Depth Estimation: A Survey," *arXiv preprint arXiv:1901.09402*.

Cao ZL, Zhong-Hong Y and Hong W (2015), "Summary of Binocular Stereo Vision Matching Technology," *Journal of Chongqing University of Technology* (*Natural Science*), **29**(2): 70–75.

Chang A, Dai A, Funkhouser T, Halber M, Niessner M, Savva M, Song S, Zeng A and Zhang Y (2017), "Matterport3D: Learning from RGB-D Data in Indoor Environments," *International Conference on 3D Vision* (*3DV*).

Cheng Y, Cai R, Li Z, Zhao X and Huang K (2017), "Locality-Sensitive Deconvolution Networks with Gated Fusion for RGB-D Indoor Semantic Segmentation," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3029–3037.

Deng J, Dong W, Socher R, Li LJ, Li K and Fei-Fei L (2009), "ImageNet: A Large-Scale Hierarchical Image Database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 248–255.

Elkins EB (2020), "Simulating Destruction Effects in SideFX Houdini," *Undergraduate Honors Theses*, Paper 524. https://du.etsu.edu/honors/524

Gao Y and Khalid MM (2022), "Deep Learning Visual Interpretation of Structural Damage Images," *Journal of Building Engineering*, p. 105144.

Groenendijk R (2020), "On the Benefit of Adversarial Training for Monocular Depth Estimation," *Computer Vision and Image Understanding*, **190**, p. 102848.

Gunasekar K, Qiang Q and Yezhou Y (2020), "Low to High Dimensional Modality Hallucination Using Aggregated Fields of View," *IEEE Robotics and Automation Letters*, **5**(2): 1983–1990.

Hazirbas C, Ma L, Domokos C and Cremers D (2016), "FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture," *Asian Conference on Computer Vision. Springer*, 213–228.

Hoffman J, Saurabh G, and Trevor D (2016), "Learning with Side Information Through Modality Hallucination," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 826–834.

Huang G, Liu Z, Van Der Maaten L and Weinberger KQ (2017), "Densely Connected Convolutional Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.

Kim H, Lee S, Ahn E, Shin M and Sim SH (2021), "Crack Identification Method for Concrete Structures Considering Angle of View Using RGB-D Camera-Based Sensor Fusion," *Structural Health Monitoring*, **20**(2): 500–512.

Kumar ACS, Suchendra MB and Mukta P (2018), "Monocular Depth Prediction Using Generative Adversarial Networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 300–308.

Kwak DH and Lee SH (2020), "A Novel Method for Estimating Monocular Depth Using Cycle GAN and Segmentation," *Sensors*, **20**(9): 2567.

Lazaros N, Georgios CS and Antonios G (2008), "Review of Stereo Vision Algorithms: From Software to Hardware," *International Journal of Optomechatronics*, **2**(4): 435–462.

Le L, Andrew P and Martha W (2018), "Supervised Autoencoders: Improving Generalization Performance with Unsupervised Regularizers," *Advances in Neural Information Processing Systems*, **31**.

Lore KG, Reddy K, Giering M and Bernal EA (2018). "Generative Adversarial Networks for Depth Map Estimation from RGB Video," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 1177–1185.

Mondal TG (2021), "Development of Multimodal Fusion-Based Visual Data Analytics for Robotic Inspection and Condition Assessment," *PhD Thesis*, Purdue University, USA.

Mondal TG and Jahanshahi MR (2020), "Autonomous Vision-Based Damage Chronology for Spatiotemporal Condition Assessment of Civil Infrastructure Using Unmanned Aerial Vehicle," *Smart Structures and Systems, An International Journal*, **25**(6): 733–749.

Mondal TG and Jahanshahi MR (2022), "Applications of Depth Sensing for Advanced Structural Condition Assessment in Smart Cities," *The Rise of Smart Cities*, Elsevier, 305–318.

Mondal TG, Jahanshahi MR, Wu RT and Wu ZY (2020), "Deep Learning-Based Multi-Class Damage Detection for Autonomous Post-Disaster Reconnaissance," *Structural Control and Health Monitoring*, **27**(4): e2507.

NCREE (2016), 2016 Taiwan Meinong Earthquake. https://datacenterhub.org/deedsdv/publications/view/534.

Ophoff T, Kristof VB and Toon G (2018), "Improving Real-Time Pedestrian Detectors with RGB+ Depth Fusion," *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance* (*AVSS*), IEEE, 1–6.

Özyeşil O, Voroninski V, Basri R and Singer A (2017), "A Survey of Structure from Motion," *Acta Numerica*, **26**: 305–364.

Park SJ, Ki-Sang H and Seungyong L (2017), "RDFNet: RGB-D Multi-Level Residual Feature Fusion for Indoor Semantic Segmentation," *Proceedings of the IEEE International Conference on Computer Vision*, 4980–4989.

Schonberger JL and Jan-Michael F (2016), "Structure-from-Motion Revisited," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Schwarz M (2018), "RGB-D Object Detection and Semantic Segmentation for Autonomous Manipulation in Clutter," *The International Journal of Robotics Research*, **37**(4-5): 437–451.

Shah P, Pujol S, Puranam A and Laughery L (2015), Database on Performance of Low-Rise Reinforced Concrete Buildings in the 2015 Nepal Earthquake, https://datacenterhub.org/resources/238.

Sim C, Villalobos E, Smith JP, Rojas P, Pujol S, Puranam AY and Laughery L (2016), Performance of Low-rise Reinforced Concrete Buildings in the 2016 Ecuador Earthquake, https://datacenterhub.org/resources/14160.

Simonyan K and Zisserman A (2014), “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *arXiv preprint arXiv:1409.1556*.

Tan DS, Yao CY, Ruiz Jr C and Hua KL (2019), “Single-Image Depth Inference Using Generative Adversarial Networks,” *Sensors*, **19**(7): 1708.

Ullman S (1979), “The interpretation of structure from motion,” *Proceedings of the Royal Society of London, Series B. Biological Sciences*, **203**(1153): 405–426.

Wang Z, Zhang Y, Mosalam KM, Gao Y and Huang SL (2022), “Deep Semantic Segmentation for Visual Understanding on Construction Sites,” *Computer-Aided Civil and Infrastructure Engineering*, **37**(2): 145–162.

Wu C (2011), “VisualSFM: A Visual Structure from Motion System,” http://www.cs.washington.edu/homes/ccwu/vsfm.

Xu X, Li Y, Wu G and Luo J (2017), “Multi-Modal Deep Feature Learning for RGB-D Object Detection,” *Pattern Recognition*, **72**: 300–313.

Yeum CM, Dyke SJ, Benes B, Hacker T, Ramirez J, Lund A and Pujol S (2019), “Postevent Reconnaissance Image Documentation Using Automated Classification,” *Journal of Performance of Constructed Facilities*, **33**(1): 04018103.

Zennaro S, Munaro M, Milani S, Zanuttigh P, Bernardi A, Ghidoni S and Menegatti E (2015), “Performance Evaluation of the 1st and 2nd Generation Kinect for Multimedia Applications,” *2015 IEEE International Conference on Multimedia and Expo* (*ICME*), IEEE, 1–6.

Zhao C, Sun Q, Zhang C, Tang Y and Qian F (2020), “Monocular Depth Estimation Based on Deep Learning: An overview,” *Science China Technological Sciences*, 1–16.

Zhou S and Song W (2020), “Deep Learning–Based Roadway Crack Classification with Heterogeneous Image Data Fusion,” *Structural Health Monitoring*, p. 1475921720948434.

Zhu JY, Park T, Isola P and Efros AA (2017). “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232.

Zou L and Yan L (2010), “A Method of Stereo Vision Matching Based on OpenCV,” *2010 International Conference on Audio, Language and Image Processing*, IEEE, 185–190.