Contents lists available at ScienceDirect

Results in Optics

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

hNet: Single-shot 3D shape reconstruction using structured light and h-shaped global guidance network

Hieu Nguyen[a,b], Khanh L. Ly[c], Tan Tran[d], Yuzheng Wang[e], Zhaoyang Wang[a,*]

[a] Department of Mechanical Engineering, The Catholic University of America, Washington, DC 20064, USA
[b] Neuroimaging Research Branch, National Institute on Drug Abuse, National Institutes of Health, Baltimore, MD 21224, USA
[c] Department of Biomedical Engineering, The Catholic University of America, Washington, DC 20064, USA
[d] Department of Electrical Engineering and Computer Science, The Catholic University of America, Washington, DC 20064, USA
[e] School of Mechanical Engineering, University of Jinan, Jinan, Shandong 250022, China

ARTICLE INFO

Keywords:
Three-dimensional image acquisition
Three-dimensional sensing
Single-shot imaging
Structured light
Convolutional neural network
Deep learning
Machine vision

ABSTRACT

Retrieving three-dimensional (3D) shape information from a single two-dimensional (2D) image has recently gained enormous attention in a variety of fields. In spite of recent advancements in algorithms and hardware developments, the easy-to-use characteristics and the accuracy of the 3D shape reconstruction are always of great interest. This paper presents a robust 3D shape reconstruction technique that integrates structured-light 3D imaging scheme with deep convolutional neural network (CNN) learning. The structured-light patterns facilitate the featuring process while the CNN modeling surpasses the complexity of the traditional 3D shape reconstructions. In the supervised learning pipeline, the input is either a single fringe-pattern or a single speckle-pattern image, and the output is its corresponding high-accuracy 3D shape label. Unlike the well-received autoencoder-based CNN model, a global guidance network path with multi-scale feature fusion is introduced into the CNN model to improve the accuracy of the 3D shape reconstruction. Experimental evaluations have been conducted to demonstrate the validity and robustness of the proposed technique, which provides a promising tool for ever-increasing scientific research and engineering applications.

1. Introduction

Three-dimensional (3D) imaging and shape reconstruction is a process of generating 3D geometric representation of objects, and it has a long-standing history in computer vision. Over the past decades, 3D imaging and shape reconstruction has been exploited in a wide range of applications such as object identification, medical practice, reserve engineering, quality assurance, biometric security, unmanned transportation, 3D printing, 3D modeling and animation (Sansoni et al., 2009; Remondino and El-Hakim, 2006; Ma et al., 2018). At present, prevalent 3D imaging and shape reconstruction methods include computed tomography, laser scanning, optical interferometry, holography, photogrammetry, structured-light technique, and so on (Chen et al., 2000; Kieu et al., 2014; Blais et al., 2004; Su et al., 2010; Wang et al., 2015). Among these techniques, the structured-light imaging technique offers accurate 3D shape measurements of objects in a real-time manner and is thus one of the most widely used techniques for 3D reconstruction in both academia and industry (Zhang, 2018; Geng, 2011; Nguyen et al., 2015; Wu et al., 2011). This can be seen from the fact that many consumer-grade red-green–blue and depth (RGB-D) sensors, such as Microsoft Kinect, Asus Xtion, Apple TrueDepth, and Intel RealSense sensors, are built on the structured-light technology in the last decade (Khoshelham et al., 2012; Keselman et al., 2017; iProov et al., 2018; Nguyen et al., 2017). Despite their popularity, however, broader applications of the low-cost commercial structured-light-based 3D reconstruction systems are often impeded by their inadequate accuracy. Therefore, the development of low-cost structured-light techniques capable of providing fast-speed and high-accuracy 3D imaging and shape reconstruction remains of significant interest.

In recent years, the computer technology field has experienced unprecedented evolution thanks to the breakthrough of artificial intelligence, specifically deep learning. Deep learning has been utilized to solve challenging problems in computer vision tasks, e.g., pattern recognition, image transformation, semantic segmentation, motion tracking, and scene understanding (Voulodimos et al., 2018; Ouyang et al., 2017; Doulamis et al., 2016; Lin et al., 2016; Noh et al., 2015). Prominent deep learning models and methods used in the

[*] Corresponding author.
E-mail address: wangz@cua.edu (Z. Wang).

https://doi.org/10.1016/j.rio.2021.100104
Received 20 February 2021; Revised 17 April 2021; Accepted 1 May 2021

computer vision applications include but not limited to: convolutional neural network (CNN), deep Boltzmann machine and deep belief network, recurrent neural network, long short-term memory, stacked denoising autoencoder, generative adversarial network, etc (Krizhevsky et al., 2017; Liu et al., 2015a; Salakhutdinov and Hinton, 2009; Vincent et al., 2010; Hochreiter and Schmidhuber, 1997; Goodfellow et al., 2014). In the fields of optics and mechanics where computer-vision-based techniques are adopted to carry out accurate measurements of physical quantities, the CNN-based frameworks have been successfully employed in numerous applications, including fringe analysis, phase retrieval, interferogram denoising, motion and deformation analysis, distance and depth determination, 3D shape measurement, etc. (Boukhtache et al., 2021; Nguyen et al., 2020; Lin et al., 2020; Feng et al., 2019; Yan et al., 2019; Ren et al., 2019; Nguyen et al., 2021a). In the area of 3D shape reconstruction, for instance, Jeught et al. (2019) proposed to predict the 3D height map from a single deformed fringe pattern by training a CNN model on a large collection of simulation data. Zheng et al. (2020) recently introduced an approach to build a digital twin of a fringe projection system, which can help extract 3D geometry information from a single-shot fringe image.

Deep learning has been helpful in improving the performance of 3D imaging and shape reconstruction via supervised learning process. It uses multiple network layers to progressively learn and transform the input into multiple levels of abstract representations to capture the complex structures of the input (LeCun et al., 2016; Bengio et al., 2013). By tuning the conversion model to fit a training dataset consisting of many data samples, it allows the 3D representation of an object to be acquired from a single or multiple 2D images of the object. An exemplar of deep learning implementation in the 3D shape reconstruction can be found in the work of Liu et al. (2015b), where a deep CNN model is capable of estimating the depth of a general scene from a single image. Following that, Laina et al. (2016) developed a fully convolutional architecture that is combined with residual learning to improve the estimation of the depth map from a single RGB image. Recently, the CNN has also been employed by some researchers to reconstruct the 3D geometry from a single RGB image upon conducting several refinement steps (Niu et al., 2018; Wang et al., 2018). It is noted that the existing single-shot 3D shape reconstruction techniques based on supervised learning process normally use datasets captured by commercial RGB-D sensors or generated by computer-aided design (CAD) (Silberman et al., 2011; Chang et al., 2015; Xiang et al., 2014). Nevertheless, the performance of the former approaches is usually not robust enough due to two factors: (1) inadequate features in the plain RGB input, and (2) relatively low accuracy of the ground truth. Likewise, the latter schemes of training network models with CAD simulation datasets generally do not adapt well in real-world applications.

In this paper, an innovative single-shot 3D shape reconstruction technique is introduced by taking full advantage of the structured-light and the deep CNN methods. The basic idea is to transform a single-shot image, which is either a speckle-pattern or a fringe-pattern image, into a 3D image using an end-to-end deep CNN model. Fig. 1 shows a schematic of the proposed single-shot 3D shape reconstruction system using structured-light patterns. Through projecting structured patterns onto the target of interest, additional features can be encoded into the single-image input to firmly improve the representation learning of object shapes. Importantly, a well-established structured-light technique, named fringe projection profilometry (FPP), is adopted in this work to generate high-accuracy 3D ground-truth labels to assure the reliability and practicability of the involved deep CNN model. It is noteworthy that the existing 3D reconstruction networks are mainly based on autoencoder architectures (Yu et al., 2020; Yang et al., 2017; Liang et al., 2020; Yan et al., 2018; Lu et al., 2018; Knyaz et al., 2017; Tatarchenko et al., 2019) that reconstruct the depth map solely from the finest layer. Inspired by the successful multi-level feature aggregation and multi-scale feature extraction networks (Qin et al., 2020; Wang et al., 2017; Liu et al., 2019; Zhao et al., 2017), the proposed network model aims to reconstruct the 3D shape map with higher accuracy by retrieving and integrating the features from multiple contextual levels. Because the encoder, decoder, and global guidance paths form an h-shaped schematic pattern, it is called **hNet** in this work. The proposed approach can gradually improve the 3D shape reconstruction accuracy through the h-shaped autoencoder backbone networks without adding complicated layers and modules. Moreover, it is considerably simpler, yet still robust, than the state-of-the-art 3D shape reconstruction techniques that rely on image registration and triangulation computation.

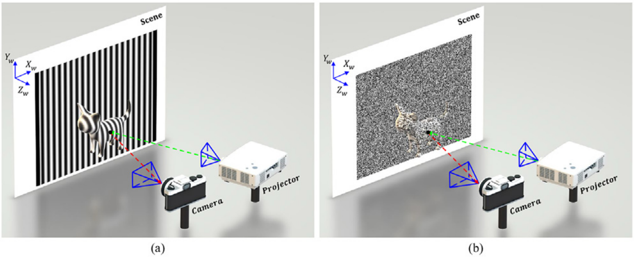

Fig. 1. Schematic of the single-shot 3D shape reconstruction system with structured-light pattern: (a) fringe pattern; (b) speckle pattern.

The remainder of this paper is organized as follows. Section 2.1 describes details of the FPP technique for generating datasets with high-accuracy ground-truth labels. Section 2.2 depicts the proposed hNet and the existing state-of-the-art UNet architectures. Section 3 demonstrates a few experimental results and presents accuracy assessment. Lastly, a discussion and a brief summary are concluded in Section 4.

2. Methodology

The purpose of the proposed technique is to transform a single structured-light image into its corresponding 3D image using deep CNNs. To prepare datasets with reliable and accurate ground-truth labels for the deep network learning, a high-accuracy FPP technique is adopted for 3D label creation, as illustrated in Fig. 2. In the generation of the required training and validation datasets, each data sample involves using a camera to capture an image of the target with a structured pattern, which is a projected speckle pattern for the case of speckle-pattern input and is a projected fringe pattern for the case of fringe-pattern input. The corresponding 3D label in pair with each input is generated by capturing a series of FPP images with phase-shifted fringe patterns. Once the training and validation datasets are ready, they can be loaded directly into the deep CNNs for subsequent learning using back-propagation optimization. After the learning process, independent test or real-application datasets can be fed into the trained artificial neural network for the purposes of assessment or application.

2.1. Fringe projection profilometry (FPP) technique for training data generation

Fig. 2(a) shows a typical setup of the FPP system which consists of a camera and a projector. The 3D shape reconstruction procedure is illustrated in Fig. 2(b), which mainly include: (1) projecting a set of fringe patterns with different spatial frequencies onto target surface, (2) capturing the distorted fringe patterns where the depth information is naturally encoded into, (3) retrieving the phase distributions from phase-shifted fringe images, and (4) computing the depth map and 3D shape. The most common automated phase-retrieval scheme

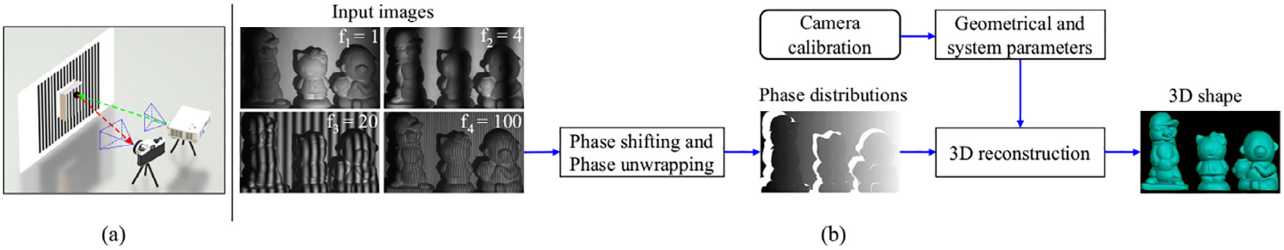

Fig. 2. Illustration of the FPP system for the 3D label generation of the datasets.

relies on using phase-shifted fringe patterns, which determines how the fringes patterns are generated and used. Here, the raw fringe patterns are evenly-spaced sinusoidal fringe patterns in vertical orientation. The fringes are digitally created using the following function (Nguyen et al., 2019; Le et al., 2018):

$$I_j^{(p)}(U,V) = I_0^{(p)}\left[1+\cos(\phi+\delta_j)\right] = I_0^{(p)}\left[1+\cos(2\pi f\frac{U}{W}+\delta_j)\right] \quad (1)$$

where $I^{(p)}$ is the intensity at pixel coordinate (U,V) in the raw pattern image; the subscript j denotes the jth phase-shifted fringe image with $j=\{1,2,\dots,m\}$, and m is the number of total phase-shift steps (e.g., $m=4$); $I_0^{(p)}$ is a constant coefficient indicating the value of intensity modulation (e.g., $I_0^{(p)}=127.5$); f is the fringe frequency, defined as the total number of fringes in the pattern; W is the width of the generated image; $\delta_j = 2\pi(j-1)/m$ is the phase-shift amount; ϕ is the fringe phase and is a function of U in the raw image.

The raw fringe patterns are projected by the projector onto the target of interest, and they are then captured by the camera as digital images. The fringes in the captured images can be described as:

$$I_j(u,v) = A(u,v) + B(u,v)\cos\left[\phi(u,v)+\delta_j\right] \quad (2)$$

where $I(u,v)$ is the intensity value at pixel coordinate (u,v) in the captured fringe image; A and B indicate the background and fringe amplitude intensities, respectively. Unlike the phase in the raw fringe patterns, the phase ϕ in the capture images is a function of both u and V.

The phase distribution of the fringes in the captured images can be determined by using a conventional phase-shifting algorithm. With the typical four-step phase-shifting scheme, the phase can be determined as:

$$\phi^w(u,v) = \arctan\frac{I_4(u,v)-I_2(u,v)}{I_1(u,v)-I_3(u,v)}, \quad (3)$$

where the superscript w denotes that the phase is wrapped since it is obtained from the inverse tangent function and is limited in a range of 0 to 2π; and the subscript numbers 1–4 represent the sequential steps of the four phase-shifted patterns in the images. Hereafter, (u,v) will be omitted from the phase distribution ϕ for simplicity.

The determination of the true phase ϕ from the wrapped one ϕ^w requires a phase-unwrapping process, which is a challenging task in practice if the application encounters geometric discontinuities such as objects of complex shapes or multiple separate objects. Accordingly, a robust scheme of using hierarchical multi-frequency phase-shifted fringe patterns is adopted. The true or unwrapped phase distributions can be calculated from the following equation (Wang et al., 2010):

$$\phi_i = \phi_i^w + \mathrm{INT}\left(\frac{\phi_{i-1}\frac{f_i}{f_{i-1}} - \phi_i^w}{2\pi}\right)2\pi \quad (4)$$

where ϕ is the unwrapped true phase; i indicates the ith fringe-frequency pattern with $i=\{2,3,\dots,n\}$, and n is the number of various fringe frequencies; INT represents the function of rounding to the nearest integer; f_i is the number of fringes in the ith projection pattern, with $f_n > f_{n-1} > \dots > f_1 = 1$; and $\phi_1 = \phi_1^w$ is automatically satisfied for $f_1 = 1$. Specifically, a multi-frequency phase-shifting scheme with $n=4$, $f_4=100$, $f_3=20$, $f_2=4$, and $f_1=1$ is employed. For each sample of accurate 3D shape reconstruction, the synchronized FPP system captures four images for each of the four frequencies, which indicates a total of 16 images.

The next step of the FPP 3D imaging technique is to calculate the height and 3D shape maps from the unwrapped phase distributions. A flexible and accurate algorithm is employed here to retrieve the 3D shape information directly from the phase distributions with the highest possible accuracy (Vo et al., 2012). The governing equation of the key coordinate z determination is (Nguyen et al., 2021b):

$$\begin{aligned}
z &= \frac{\left[\boldsymbol{p}_1\ \boldsymbol{p}_2\ \boldsymbol{p}_3\ \boldsymbol{p}_4\right]\boldsymbol{c}^{\mathsf{T}}}{\left[\boldsymbol{p}_1\ \boldsymbol{p}_2\ \boldsymbol{p}_3\ \boldsymbol{p}_4\right]\boldsymbol{d}^{\mathsf{T}}}\\
\boldsymbol{p}_1 &= \left[1\ \ \phi\ \ u\ \ u\phi\ \ v\ \ v\phi\ \ \right]\\
\boldsymbol{p}_2 &= \left[u^2\ \ u^2\phi\ \ uv\ \ uv\phi\ \ v^2\ \ v^2\phi\ \ \right]\\
\boldsymbol{p}_3 &= \left[u^3\ \ u^3\phi\ \ u^2v\ \ u^2v\phi\ \ uv^2\ \ uv^2\phi\ \ v^3\ \ v^3\phi\right]\\
\boldsymbol{p}_4 &= \left[u^4\ \ u^4\phi\ \ u^3v\ \ u^3v\phi\ \ u^2v^2\ \ u^2v^2\phi\ \ uv^3\ \ uv^3\phi\ \ v^4\ \ v^4\phi\right]\\
\boldsymbol{c} &= \left[1\ \ c_1\ \ c_2\ \ c_3\ \ \cdots\ \ c_{27}\ \ c_{28}\ \ c_{29}\right]\\
\boldsymbol{d} &= \left[d_0\ \ d_1\ \ d_2\ \ d_3\ \ \cdots\ \ d_{27}\ \ d_{28}\ \ d_{29}\right]
\end{aligned} \quad (5)$$

where z is the z-coordinate of a point in the reference or world coordinate system and the point is corresponding to a pixel (u,v) in the captured images; ϕ is the unwrapped phase obtained in Eq. (4); and $c_1 - c_{29}$ and $d_0 - d_{29}$ are constant coefficients that are related to the geometrical and other system parameters. The 59 coefficients in vectors $\boldsymbol{c}$ and $\boldsymbol{d}$ can be determined by a calibration process using a flat calibration board positioned at a number of arbitrarily locations and orientations (Vo et al., 2012).

Upon obtaining the z-coordinates or depth map of the points correlated with the pixels in the 2D image(s), the corresponding x and y coordinates can be easily calculated using the camera calibration parameters. Consequently, the two terms, 3D shape reconstruction and depth measurement, are often used interchangeably. Using the depth map (u,v,z) other than the 3D coordinates (x,y,z) in the CNN-based learning process has a notable benefit that only z is the output of interest.

The afore-described FPP technique is employed to generate the datasets with ground-truth labels, as demanded by the supervised network learning. In the data capturing and labeling process, each sample target is sequentially illuminated by the projector with 17 structured-light patterns, and 17 images are accordingly captured. The first image is a speckle-pattern image with its initial projection image pre-generated with random speckles, and the rest 16 images are the multi-frequency fringe-pattern images previously described. The 16 fringe images are utilized to determine a high-accuracy 3D depth map as the ground-truth output, while the first and last of the 17 images are chosen to serve as the inputs, i.e., a speckle-pattern image and a high-frequency fringe image. Consequently, two types of datasets are created by using two different inputs and an identical output: one is speckle-pattern dataset and the other is fringe-pattern dataset. They allow fair assessments and comparisons to identify which kind of structured-light pattern performs better for the single-shot 3D shape reconstructions.

It took about four hours to generate the entire datasets for the proposed work, and the labeling pipeline totally prepared 1200 training data samples. The datasets can be directly fed into various deep CNN models to implement the learning process, where 10% of the data samples are randomly selected to act as validation datasets to monitor the convergence of the loss function for the network learning models. In addition to preparing the training and validation datasets, the FPP system was also employed to produce 60 test data samples with 3D shape labels for each of the two types of inputs. The test datasets are independent and useful for evaluation and assessment purposes. The datasets have been made publicly available for the research community (Single-shot 3d shape reconstruction datasets, 2020).

2.2. Network architecture

The proposed hNet structure comprises of three paths: the encoder path (**E**), the decoder path (**D**), and the global guidance path (**G**). The architecture of hNet is illustrated in Fig. 3, where an “h” shape can be seen. The encoder and decoder paths are basically the autoencoder-based backbone, which is the core structure of the well-known UNet (Ronneberger et al., 2015), as shown in Fig. 4. Like the popular CNN architectures (Krizhevsky et al., 2017; Liu et al., 2015a) adopted for image classification applications, the encoder path consists of spatial convolutions with a kernel size of 3 × 3 and max-pooling layers to extract the local contextual information from the input feature maps. The feed-forward networks in the encoder path provide probability vector maps at coarse-level layers with enlarged receptive field sizes. By contrast, the decoder path with transpose and normal convolutions gradually projects the lower resolution feature maps to higher resolution layers. In the network, the symmetric concatenations between the encoder and decoder layers enable the precise transformation of data at different resolution levels.

A typical autoencoder-based network only reconstructs the output feature map with fine-level contextual information where the resolution of the output is the same as that of the input, such as D_4 shown in Fig. 3. However, it is believed that the coarse-level features at

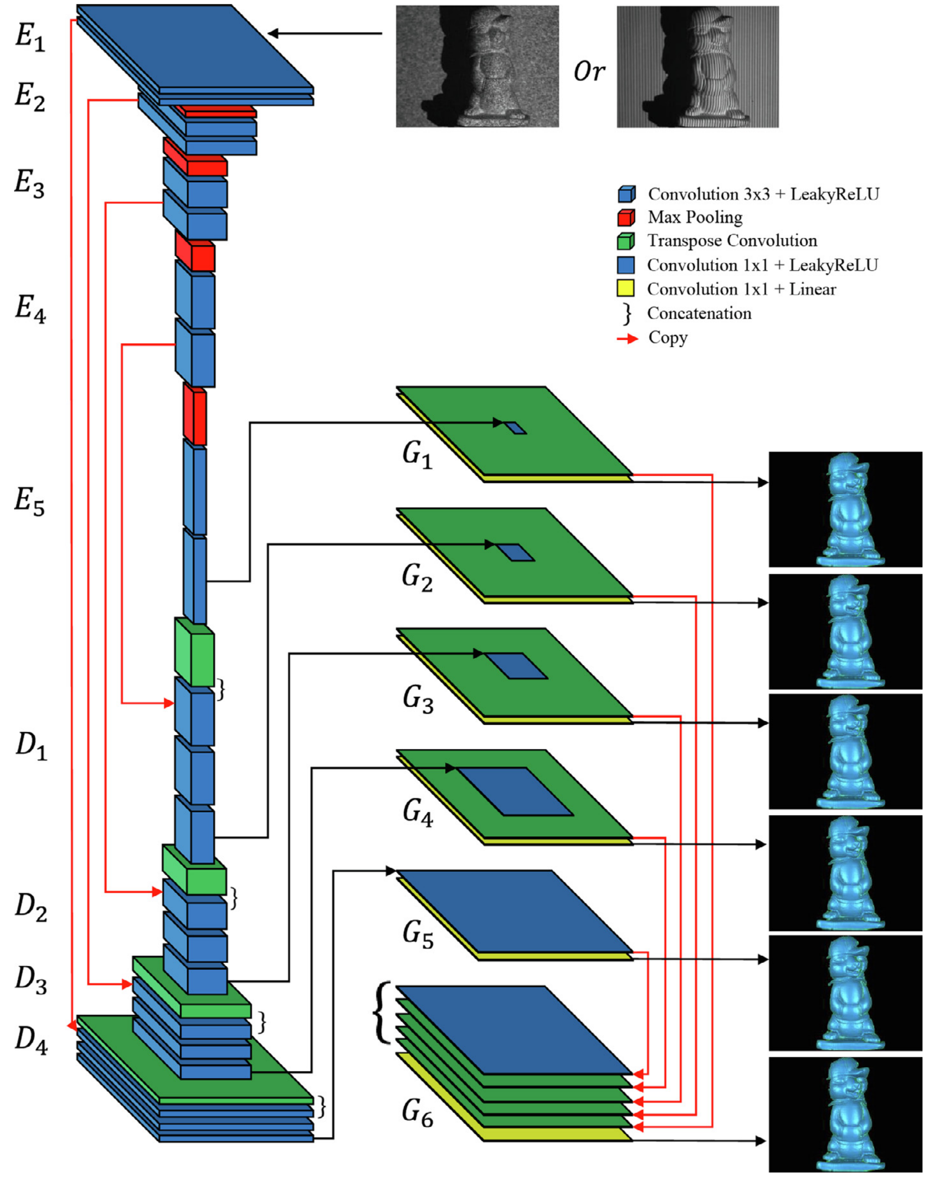

Fig. 3. Architecture of the proposed network.

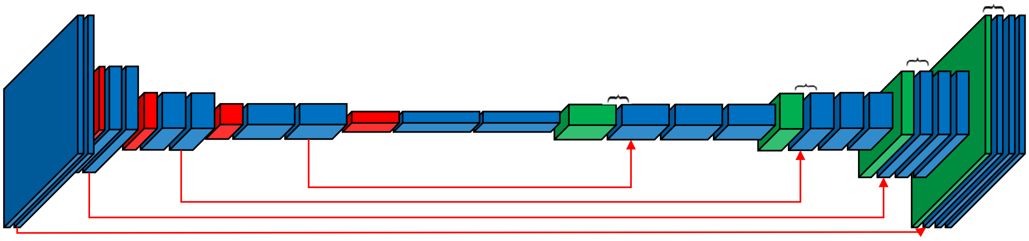

Fig. 4. Architecture of the UNet network selected for comparison.

low-resolution feature maps E_5, D_1, D_2, and D_3 are also helpful for improving the accuracy of the output. Because both global and local contextual information from the feature maps at the coarse-level and fine-level layers are essential in the 3D shape reconstruction, this work proposes using a global guidance path to provide extra global or coarse features to the highest fine-level feature map in G_6. The 1×1 convolution and transpose convolution with a channel depth of 1 have been applied to each resolution in E_5, D_1, D_2, D_3 and D_4 to leverage the receptive field size to 1 as well as upsample the resolution to the size of the input. Finally, a 1×1 convolution with linear activation is applied at each block in $\mathbf{G}$ to bring the information of vector feature map to the corresponding 3D label. Since the decoder block D_4 is of the same resolution as the input, there is no need to use transpose convolution in G_5 for further upsampling. In general, the proposed network has a total of 6 outputs from 6 different blocks in the $\mathbf{G}$ path where the output 3D labels are identical.

The global guidance path contributes to the 3D reconstruction accuracy in two ways. First, the feature maps from coarse-level to fine-level, E_5–D_4, are optimized and trained precisely by forcing G_1–G_5 to predict the exact high-accuracy 3D labels. Second, the explicit feature maps in G_1–G_5 have been concatenated and fused in G_6 for the final prediction. The coarse-level and fine-level feature maps from $G_1 - G_5$ globally support the output feature map in G_6, and together all the $\mathbf{G}$ blocks generate a global guidance network. In hNet, only the prediction from G_6 is considered for the final 3D shape prediction and subsequent performance assessments. Since hNet contains six outputs, the total loss of the training process can be determined as (Xie et al., 2015):

$$\mathbb{L} = \omega_c l_c + \sum_{n=1}^{N} \omega_n l_n \tag{6}$$

where l_c is the loss of the final concatenation layer in G_6; ,l_n denotes the loss for other output layers in $\mathbf{G}$ with $n = \{1, 2, \ldots, N\}$ and $N = 5$; ,w_c and w_1–w_5 are the weights contributed to each corresponding loss.

In the learning process, a back-propagation algorithm is adopted to minimize the learning loss as well as update the training parameters after each epoch. The back-propagation is performed based on the chain architecture, where each hidden layer in the network is defined as (Goodfellow et al., 2016):

$$\mathbf{h} = g(\mathbf{W}^{\intercal}\mathbf{x} + \mathbf{b}) \tag{7}$$

where $\mathbf{x}$ is a vector of the input; $\mathbf{h}$ is the output vector; function g is an activation function; the weight parameters $\mathbf{W}$ in a matrix form and the bias parameters $\mathbf{b}$ in a vector form are optimized by the training process.

The data format of the training, validation, and test datasets is a four-dimensional (4D) tensor of size $s \times h \times w \times c$, where s is the number of the data samples; h and w denote the height and width of each image, respectively; c is the channel depth of each data, with $c = 1$ for gray-scale images and $c = 3$ for color RGB images. Besides the normal convolutions, max-pooling operations, and transpose convolutions, a dropout regularization process is applied to the desired layer in the network where the dropout rate is set to 0.2. This means that 20% of the inputs will be randomly excluded from each update cycle to prevent overfitting issue. Furthermore, a nonlinear activation function name leaky rectified linear unit (LeakyReLU) is employed to cope with the zero gradient issue in the rectified linear unit (ReLU). The LeakyReLU activation function can be described as:

$$\mathbf{h}' = g(\mathbf{W}^{\intercal}\mathbf{x} + \mathbf{b}) = \begin{cases} \mathbf{W}^{\intercal}\mathbf{x} + \mathbf{b} & \mathbf{W}^{\intercal}\mathbf{x} + \mathbf{b} > 0 \\ \alpha(\mathbf{W}^{\intercal}\mathbf{x} + \mathbf{b}) & \mathbf{W}^{\intercal}\mathbf{x} + \mathbf{b} \leqslant 0 \end{cases} \tag{8}$$

where α is a negative slope coefficient.

Compared with our preceding work (Nguyen et al., 2020), the work proposed in this paper has two essential improvements. Unlike the previous autoencoder-based networks, the novel network has a h-shaped global guidance path to help improve the accuracy of the 3D shape reconstruction through extracting additional information from the inputs. Another substantial difference is that the new datasets contain both fringe-pattern images and speckle-pattern images (Single-shot 3d shape reconstruction datasets, 2020). Consequently, the new datasets allow the proposed network to use not only fringe-pattern images but also speckle-pattern images, which are two most widely used structured-light patterns in practice.

In contrast to the state-of-the-art 3D shape reconstruction techniques which are based on profound understanding of the system geometrics and sophisticated algorithms, the proposed technique relies on numerous parameters and hyperparameters in the artificial neural networks, which are automatically decided by going through an end-to-end training process.

3. Experiments and results

Several experiments have been carried out to demonstrate the capability and robustness of the proposed deep-learning approach for 3D shape reconstruction. For computation efficiency, a dedicated desktop computer for deep learning with an Intel Core i9-9900 K CPU, a 64 GB RAM, and an NVIDIA GeForce RTX 2080 Ti graphics card is used. The NVIDIA parallel computing platform named CUDA (compute unified device architecture) and deep learning library named cuDNN (CUDA deep neural network) as well as the Google TensorFlow machine learning library are installed on the computer to facilitate the training process of the end-to-end neural network models. The TensorFlow has application programming interfaces (APIs) available in Python and several other languages, and the one adopted in the proposed work is the Keras deep learning API running on top of the TensorFlow.

In the system setup, the projector and camera are arbitrarily arranged to form a generalized layout. The system calibration is then completed to acquire the camera intrinsic and extrinsic parameters as well as the coefficients presented in Eq. (5). The working field of view (FOV) is approximately 150 mm, and the nominal distance from the camera to the objects is around 1.2 m.

3.1. Hyperparameter configuration

The supervised deep learning is an iteration process to find the optimal network parameters so that the output results have the best matching with the output labels of the input training dataset. The iteration uses a back-propagation algorithm and the convergence criterion is the total-loss function defined in Eq. (6). In addition to the key model parameters to be learned from dataset, there are configuration hyperparameters that should be specified manually. In this work, a total of 200 epochs with a mini-batch size of 2 is applied to the learning network. Adam optimizer (Kingma et al., 2015), a stochastic gradient descent method that is based on an adaptive estimation of first-order and second-order moments, is employed in the learning. Instead of using a fixed learning rate for the Adam optimizer, a step decay schedule is employed to gradually reduce the learning rate from its initial value (e.g. 0.0001) as the epoch number increases. Besides the learning rate control, two common callback schemes, namely LambdaCallback and ModelCheckpoint in the Keras library, are adopted to monitor and save the intermediate results of the learning. LambdaCallback helps to save the 3D shape reconstruction with the updated parameters, whereas ModelCheckpoint is set to save the best-optimized model.

A loss function of root mean squared error (RMSE) is used to monitor the optimization process. It is noted that whether the model is updated or not is based on the loss function of the validation dataset. In addition, another error, mean relative error (MRE), is also included in this work for performance assessment and comparison between the prediction results and 3D ground-truth labels. These two errors are determined as:

- Mean relative error: $\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{z}_i - z_i}{d_i}\right)$
- Root mean squared error: $\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{z}_i - z_i\right)^2}$

where N is the number of valid points; $\hat{z}_i$ and z_i are the predicted depth and the labeled ground-truth depth at the ith point, respectively; and d_i is the distance from the camera to the point of interest.

3.2. Evaluation on speckle-pattern datasets

The performances of the proposed hNet and the existing UNet models are first evaluated with speckle-pattern datasets. Shown in Figs. 5 (a) and (b) are the MRE and RMSE plots obtained during the learning process, respectively; and the values displayed in the figure are gained from the last epoch. Fig. 5(c) shows three representative results, each of which contains a plain image of the object, the speckle-input image, the ground-truth 3D shape, and the 3D shapes reconstructed by UNet and hNet, respectively. Overall, both networks were able to generate the approximate 3D shapes of the given objects with acceptable accuracy. Particularly, the 3D shapes reconstructed by hNet exhibit smoother surfaces with more details than those generated by UNet. The improvement is due to the coarse and fine features obtained from the low-level feature maps in the global guidance path.

Details of a few relevant performance data are listed in Table 1, where the MRE and RMSE values are associated with the most optimized models. The table reveals that the two networks require a nearly equal number of parameters, memory size for model, and floating-point operations (FLOPs) for the learning task. The hNet model needs slightly larger values because of the introduction of the guidance path. It is evident from the table that the MRE and RMSE values of the hNet model are lower than those of the UNet model for the training and validation datasets, as well as the test dataset that has never been used in training. This suggests that hNet can predict and reconstruct 3D shapes from speckle-pattern images more accurately than UNet.

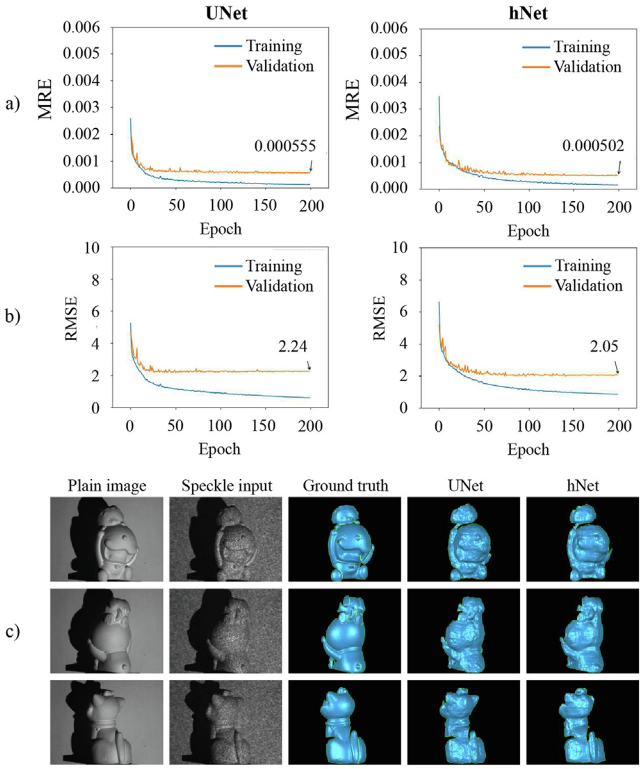

Fig. 5. 3D reconstruction results of three representative speckle-pattern inputs.

Table 1
Performance comparison of hNet and UNet with single speckle image as input.

Model		UNet	hNet
Number of parameters		8,629,921	8,632,261
Model memory size (MB)		103.7	103.9
Floating-point operations		119,217,254,408	119,253,606,016
Training	MRE	2.495×10^{-4}	2.146×10^{-4}
	RMSE	1.079	1.004
Validation	MRE	5.752×10^{-4}	5.197×10^{-4}
	RMSE	2.17	2.018
Test	MRE	6.389×10^{-4}	5.742×10^{-4}
	RMSE	2.203	2.077

3.3. Evaluation on fringe-pattern datasets

Similarly, the performance assessments have been conducted on the fringe-pattern datasets using hNet and UNet. It is important to point out that each network must be trained separately for the fringe-image input and the speckle-image input, even though the network structure remains the same and the structured-light system is geometrically unchanged. The reason is that the 3D shape information is transformed and extracted differently by the network for different structured-light patterns. Furthermore, if the geometrical configuration of a structured-light system changes, the network must be trained again with updated datasets to ensure reliable shape reconstructions.

Fig. 6 shows the MRE and RMSE plots acquired during the learning process as well as the results of the same three representative objects presented previously. Like the speckle-image input, the fringe-image input works well with both hNet and UNet. The hNet model is capable of retrieving the 3D representations of objects with higher accuracy compared with UNet. Again, the reason is that UNet reconstructs the

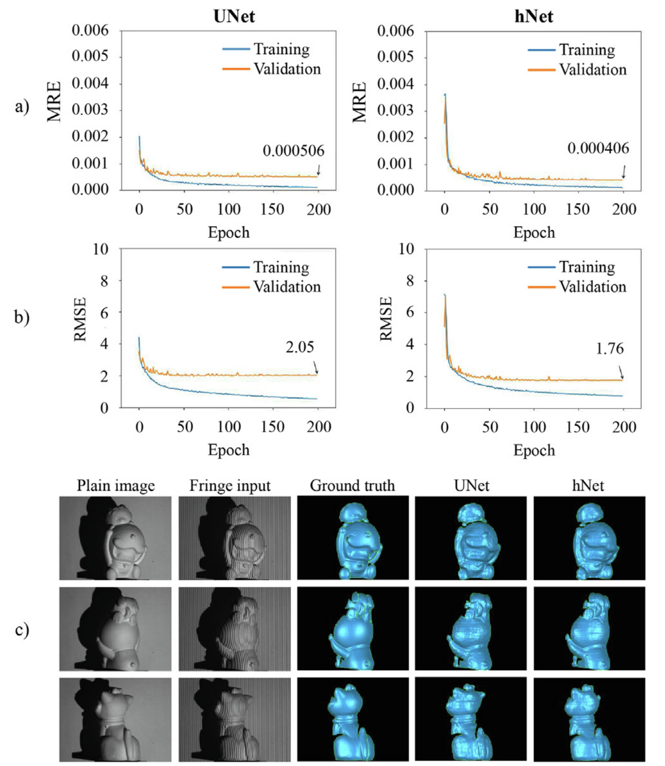

Fig. 6. 3D reconstruction results of three representative fringe-pattern inputs.

Table 2
Performance comparison of hNet and UNet with single fringe image as input.

Model		UNet	hNet
Number of parameters		8,629,921	8,632,261
Model memory size (MB)		103.7	103.9
Floating-point operations		119,217,254,408	119,253,606,016
Training	MRE	1.673×10^{-4}	1.617×10^{-4}
	RMSE	0.833	0.825
Validation	MRE	5.092×10^{-4}	4.196×10^{-4}
	RMSE	1.963	1.724
Test	MRE	6.4×10^{-4}	5.329×10^{-4}
	RMSE	2.135	1.861

3D shape from the final layer only, whereas hNet uses a total of six intermediate outputs to refine the final output.

Table 2 shows results of the selected performance data identical to those in Table 1, and same conclusion can be drawn. A comparison of the data listed in Tables 1 and 2 indicates that the 3D reconstruction performance from using the single-shot fringe patterns is better than that from using the single-shot speckle patterns. This is also evident from the visualization of the 3D reconstructions exhibited in Figs. 5 (c) and 6(c), where the 3D reconstruction results from the fringe-pattern images have finer surface and less noise than the ones reconstructed from the speckle-pattern images.

In order to show the gradual improvement of the six outputs in the guidance path, the RMSE plots of the $G_1 - G_6$ outputs documented during the learning process and the corresponding 3D outputs of the three representative test samples are demonstrated in Fig. 7. Although each group of the six outputs originate from an identical input with an identical label, they clearly have different accuracies, and the output losses gradually reduce in the order from block G_1 to block G_6.

3.4. 3D reconstruction of multiple separate objects

A third experiment has been conducted to validate the capability of the proposed deep learning approach for 3D shape reconstructions of multiple separate objects in the scene of view. Measuring the 3D shapes of multiple objects that are presented with geometric discontinuities is a challenging work for a typical single-shot fringe-projection-based method because the fringe orders among the objects are discontinuous. As observed from the results shown in Fig. 8, the discontinuity of fringe orders does not pose a problem to the 3D reconstruction of the proposed approach owing to the fact that it studies the 3D representations directly from the fringe-pattern features with a large number of training samples. Since a single-shot fringe-pattern image does not contain true fringe-order information, fringe orders are never used in the learning process. The capability of coping well with multiple separate objects indicates that the proposed hNet approach is superior to the conventional single-shot fringe projection techniques which highly rely on sophisticated phase calculation process.

3.5. Extended application to salient object detection

Salient object detection (SOD) is a task to identify and to segment the most visually attractive objects appeared in images. The capability of reconstructing 3D shape from a single image makes the proposed hNet architecture suitable for the SOD applications as well. To demonstrate the validity of the spin-off capability, two common benchmark datasets, Extended Complex Scene Saliency Dataset (ECSSD) and Frequency-tuned (FT) dataset, are chosen to evaluate the proposed network. The ECSSD comprises of 1000 structurally complicated images with large foreground objects (Yan et al., 2013; Shi et al., 2016), whereas the FT dataset consists of 1000 natural images with complicated foreground objects and background that are favorable for object detection and segmentation evaluations (Li et al., 2014; Achanta et al., 2009).

The proposed hNet architecture remains the same except some minor changes are made to accommodate the SOD applications. Since both datasets contain images of various image sizes, a data augmentation scheme has been adopted to resize the images and their corresponding ground-truth masks to the same resolution. The channel c in the 4D tensor of $s \times h \times w \times c$ size is set to 3 because the inputs are RGB images. The last activation layers from hNet is changed to sigmoid and the loss function is modified as binary cross entropy to adapt with the binary ground-truth masks, where the value of each pixel is either 0 or 1. Fig. 9 displays some qualitative results, where the first image in each row is the input image, the following image is the corresponding ground-truth mask, and the last image is the saliency map reconstructed by hNet. It can be clearly seen that the salient objects in the RGB images have been segmented correctly and close to the ground truth even in the most challenging circumstances despite some minor errors. The proposed hNet performs particularly well on the dataset if there are not significant intensity variations in the images. It is noted that the proposed hNet scheme uses a quick wrap-up tuning of 200 epochs and spends about 5 training hours for the presented work. Because of the scope of this paper, further study of the spin-off application to SOD is not yet conducted. Nonetheless, it is worth conducting a further more rigorous investigation to improve the segmentation accuracy in future work.

4. Discussions and conclusions

In this paper, an innovative single-shot 3D shape reconstruction technique integrating structured-light 3D imaging and CNN deep learning is presented. The proposed hNet approach takes full advantage of the structured-light technique and the CNN modeling, where the structured light helps facilitate the transformation of shape infor-

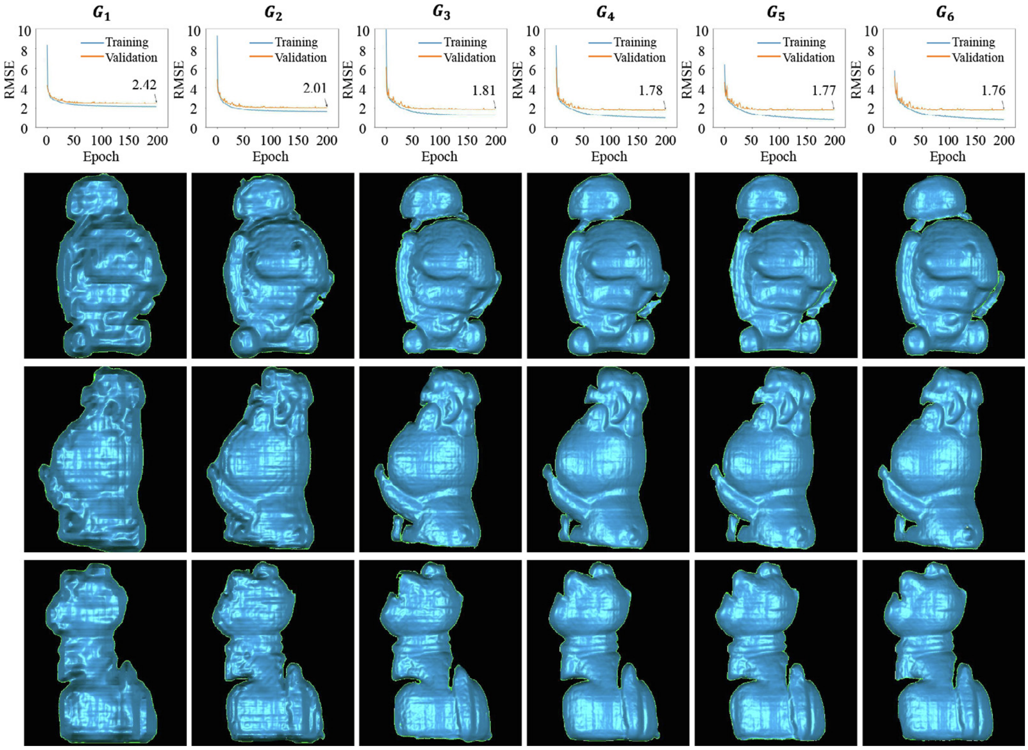

Fig. 7. 3D output comparisons of the six coarse-to-fine blocks in the global guidance path of the proposed hNet.

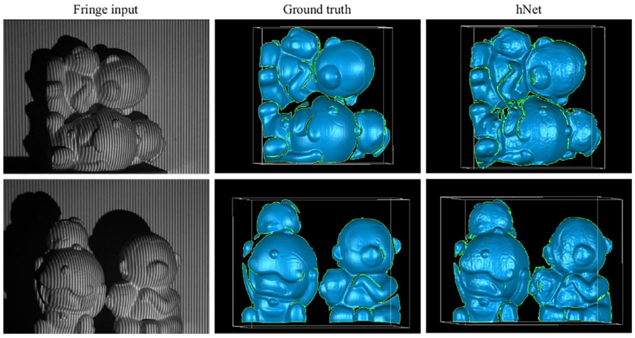

Fig. 8. 3D reconstruction results of multiple separate objects.

mation and the CNN model reconstructs the 3D shape from a single image using an end-to-end deep network. In addition, the structured-light technique provides learning datasets with high-accuracy ground-truth labels. By introducing a global guidance path to the artificial neural network, the proposed hNet technique achieves better performance than the widely-used autoencoder-based CNN model, the U-shaped UNet. Two types of structured-light patterns are investigated, and it is shown that the fringe pattern performs better than the speckle pattern. Further improvement on the proposed hNet technique may be made by multiple pipelines, including but not limited to enlarging datasets, accurately tuning hyperparameters, and exploring enhanced networks.

Currently, the measurement accuracy of the proposed technique is inferior to the state-of-the-art conventional 3D imaging and shape

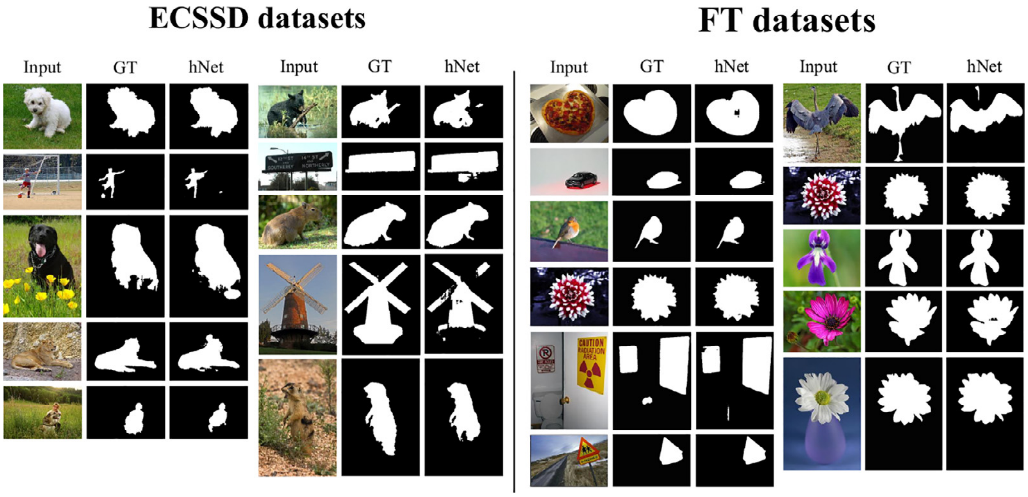

Fig. 9. Results of application to saliency map reconstruction.

reconstruction techniques. Nevertheless, the novel approach possesses an unprecedented behavior of vital importance: it requires only a single image. It is well known that in traditional theory, it is impossible to cope with geometric discontinuities in 3D shape measurements with a single-shot high-frequency fringe image. With artificial-intelligence approach, the discontinuity problem becomes solvable through deep learning from labeled datasets. Such unreasonable robustness and effectiveness have been observed in a variety of fields (Goecks et al., 2020; Sejnowski, 2020; Høye et al., 2021). It is expected that upon successful future work, the novel technique would provide very promising 3D imaging and shape-measurement solutions in innumerable scientific research and engineering applications. Moreover, the successful extension of the proposed hNet to saliency object detection demonstrates its impact in a wider range of practices.

Funding

This research received no external funding.

Declaration of Competing Interest

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

CRediT authorship contribution statement

Hieu Nguyen: Conceptualization, Methodology, formal analysis, Software, Writing - original draft. **Khanh L. Ly:** Methodology, Data curation, Investigation. **Tan Tran:** Software, Visualization. **Yuzheng Wang:** Validation, Writing - review & editing. **Zhaoyang Wang:** Conceptualization, Supervision, Writing - review & editing.

References

Achanta, R., Hemami, S., Estrada, F., Susstrunk, S., 2009. Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1597–1604. https://doi.org/10.1109/CVPR.2009.5206596.

Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE Trans. Patt. Anal. Mach. Intell. 35, 1798–1828. https://doi.org/10.1109/TPAMI.2013.50.

Blais, F., 2004. Review of 20 years of range sensor development. J. Electron. Imaging 13, 231–243. https://doi.org/10.1117/1.1631921.

Boukhtache, S., Abdelouahab, K., Berry, F., Blaysat, B., Grédiac, M., Sur, F., 2021. When deep learning meets digital image correlation. Opt. Lasers Eng. 136,. https://doi.org/10.1016/j.optlaseng.2020.106308 106308.

Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F. 2015. Shapenet: An information-rich 3d model repository, arXiv:1512.03012 [cs.GR]..

Chen, F., Brown, G.M., Song, M., 2000. Overview of three-dimensional shape measurement using optical methods. Opt. Eng. 39, 10–22. https://doi.org/10.1117/1.602438.

Doulamis, N., Voulodimos, A., 2016. Fast-mdl: Fast adaptive supervised training of multi-layered deep learning models for consistent object tracking and classification. In: IEEE International Conference on Imaging Systems and Techniques (IST), pp. 318–323. https://doi.org/10.1109/IST.2016.7738244.

Feng, S., Chen, Q., Gu, G., Tao, T., Zhang, L., Hu, Y., Yin, W., Zuo, C., 2019. Fringe pattern analysis using deep learning. Adv. Photonics 1, (2). https://doi.org/10.1117/1.AP.1.2.025001 025001.

Geng, J., 2011. Structured-light 3d surface imaging: a tutorial. Adv. Opt. Photon. 3, 128–160. https://doi.org/10.1364/AOP.3.000128.

Goecks, J., Jalili, V., Heiser, L., Gray, J., 2020. How machine learning will transform biomedicine. Cell 181, 92–101. https://doi.org/10.1016/j.cell.2020.03.022.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: International Conference on Neural Information Processing Systems (NIPS), Vol. 2, pp. 2672–2680. https://doi.org/10.5555/2969033.2969125..

Goodfellow, I., Bengio, Y., Courville, A. 2016. Deep Learning, The MIT Press, Cambridge, Massachusetts. doi: https://mitpress.mit.edu/books/deep-learning..

Hochreiter, S., Schmidhuber, J. 1997. Long short-term memory, Neur. Comp. 9, 1735–178. https://doi.org/10.1162/neco.1997.9.8.1735..

Høye, T., Ärje, J., Bjerge, K., Hansen, O., Iosifidis, A., Leese, F., Mann, H., Meissner, K., Melvad, C., Raitoharju, J. 2021. Deep learning and computer vision will transform entomology, in: Proceedings of the National Academy of Sciences, Vol. 118, p. e2002545117. https://doi.org/10.1073/pnas.2002545117..

iProov, A.B., 2018. Facing the future: the impact of apple faceid. Biom. Tech. Today 1, 5–7. https://doi.org/10.1016/S0969-4765(18)30010-9.

Jeught, S., Dirckx, J., 2019. Deep neural networks for single shot structured light profilometry. Opt. Express 27, 17091–17101. https://doi.org/10.1364/OE.27.017091.

Keselman, L., Woodfill, J.I., Grunnet-Jepsen, A., Bhowmik, A. 2017. Intel(r) realsense (tm) stereoscopic depth cameras, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1267–1276. https://doi.org/10.1109/CVPRW.2017.167..

Khoshelham, K., Elberink, S.O., 2012. Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors 12, 1437–1454. https://doi.org/10.3390/s120201437.

Kieu, H., Pan, T., Wang, Z., Le, M., Nguyen, H., Vo, M., 2014. Accurate 3d shape measurement of multiple separate objects with stereo vision. Meas. Sci. Tech. 25,. https://doi.org/10.1088/0957-0233/25/3/035401 035401.

Kingma, D., Ba, J. 2015. A method for stochastic optimization, International Conference on Learning Representations (ICLR) https://doi.org/10.1088/0957-0233/25/3/035401..

Knyaz, V.A., Vygolov, O., Kniaz, V.V., Vizilter, Y., Gorbatsevich, V., Luhmann, T., Conen, N., 2017. Deep learning of convolutional auto-encoder for image matching and 3d object reconstruction in the infrared range. In: IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 2155–2164. https://doi.org/10.1109/ICCVW.2017.252.

Krizhevsky, A., Sutskever, I., Hinton, G. 2017. Imagenet classification with deep convolutional neural networks, in: Communications of the ACM, Vol. 60, pp. 84–90. https://doi.org/10.1145/3065386..

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 239–248.

LeCun, Y., Bengio, Y., Hinton, G., 2016. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Le, H., Nguyen, H., Wang, Z., Opfermann, J., Leonard, S., Krieger, A., Kang, J., 2018. Demonstration of a laparoscopic structured-illumination three-dimensional imaging system for guiding reconstructive bowel anastomosis. J. Biomed. Opt. 23, 1–10. https://doi.org/10.1117/1.JBO.23.5.056009.

Liang, J., Zhang, J., Shao, J., Song, B., Yao, B., Liang, R., 2020. Deep convolutional neural network phase unwrapping for fringe projection 3d imaging. Sensors 20, 3691. https://doi.org/10.3390/s20133691.

Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L., 2014. The secrets of salient object segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 280–287. https://doi.org/10.1109/CVPR.2014.43.

Lin, L., Wang, K., Zuo, W., Wang, M., Luo, J., Zhang, L., 2016. A deep structured model with radius-margin bound for 3d human activity recognition. Int. J. Comput. Vis. 218, 256–273. https://doi.org/10.1007/s11263-015-0876-z.

Lin, B., Fu, S., Zhang, C., Wang, F., Li, Y., 2020. Optical fringe patterns filtering based on multi-stage convolution neural network. Opt. Lasers Eng. 126,. https://doi.org/10.1016/j.optlaseng.2019.105853 105853.

Liu, S., Deng, W., 2015a. Very deep convolutional neural network based image classification using small training sample size. In: IAPR Asian Conference on Pattern Recognition (ACPR), pp. 730–734. https://doi.org/10.1109/ACPR.2015.7486599.

Liu, F., Shen, C., Lin, G., 2015b. Deep convolutional neural fields for depth estimation from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5162–5170.

Liu, J., Hou, Q., Cheng, M., Feng, J., Jiang, J., 2019. A simple pooling-based design for real-time salient object detection. In: IEEE International Conference on Computer Vision (ICCV), pp. 3912–3921. https://doi.org/10.1109/CVPR.2019.00404.

Lu, L., Qi, L., Luo, Y., Jiao, H., Dong, J., 2018. Three-dimensional reconstruction from single image base on combination of cnn and multi-spectral photometric stereo. Sensors 18, 764. https://doi.org/10.3390/s18030764.

Ma, Z., Liu, S., 2018. A review of 3d reconstruction techniques in civil engineering and their applications. Adv. Eng. Informat. 37, 163–174. https://doi.org/10.1016/j.aei.2018.05.005.

Nguyen, H., Nguyen, D., Wang, Z., Kieu, H., Le, M., 2015. Real-time, high-accuracy 3d imaging and shape measurement. Appl. Opt. 54, A9–A17. https://doi.org/10.1364/AO.54.0000A9.

Nguyen, H., Wang, Z., Jones, P., Zhao, B., 2017. 3d shape, deformation, and vibration measurements using infrared kinect sensors and digital image correlation. Appl. Opt. 56, 9030–9037. https://doi.org/10.1364/AO.56.009030.

Nguyen, H., Dunne, N., Li, H., Wang, Y., Wang, Z., 2019. Real-time 3d shape measurement using 3lcd projection and deep machine learning. Appl. Opt. 58, 7100–7109. https://doi.org/10.1364/AO.58.007100.

Nguyen, H., Wang, Y., Wang, Z., 2020. Single-shot 3d shape reconstruction using structured light and deep convolutional neural networks. Sensors 20, 3718. https://doi.org/10.3390/s20133718.

Nguyen, H., Tran, T., Wang, Y., Wang, Z., 2021a. Three-dimensional shape reconstruction from single-shot speckle image using deep convolutional neural networks. Opt. Lasers Eng. 143,. https://doi.org/10.1016/j.optlaseng.2021.106639 106639.

Nguyen, H., Liang, J., Wang, Y., Wang, Z. Accuracy assessment of fringe projection profilometry and digital image correlation techniques for three-dimensional shape measurements, JPhys Photon. https://doi.org/10.1088/2515-7647/abcbe4..

Niu, C., Li, J., Xu, K., 2018. Im2struct: Recovering 3d shape structure from a single rgb image. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4521–4529. https://doi.org/10.1109/CVPR.2018.00475.

Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. In: IEEE International Conference on Computer Vision (ICCV), pp. 1520–1528. https://doi.org/10.1109/ICCV.2015.178.

Ouyang, W., Zeng, X., Wang, X., Qiu, S., Luo, P., Tian, Y., Li, H., Yang, S., Wang, Z., Li, H., Wang, K., Yan, J., Loy, C., Tang, X., 2017. Deepid-net:object detection with deformable part based convolutional neural networks. IEEE Trans. Patt. Anal. Mach. Intell. 39, 1320–1334. https://doi.org/10.1109/TPAMI.2016.2587642.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M. 2020. u^2-net: Going deeper with nested u-structure for salient object detection, Patt. Recogn. 106. https://doi.org/10.1016/j.patcog.2020.107404..

Remondino, F., El-Hakim, S., 2006. Image-based 3d modelling: A review. Photogram. Rec. 21, 269–291. https://doi.org/10.1111/j.1477-9730.2006.00383.x.

Ren, Z., So, H., Lam, E., 2019. Fringe pattern improvement and super-resolution using deep learning in digital holography. IEEE Trans. Industr. Inform. 15, 6179–6186. https://doi.org/10.1109/TII.2019.2913853.

Ronneberger, O., Fischer, P., Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, Vol. 9351, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28..

Salakhutdinov, R., and Hinton, G. 2009. Deep boltzmann machines, in: Conference on Artificial Intelligence and Statistics (PMLR), Vol. 5, pp. 448–455. doi:10.1109/ACPR.2015.7486599..

Sansoni, G., Trebeschi, M., Docchio, F., 2009. State-of-the-art and applications of 3d imaging sensors in industry, cultural heritage, medicine, and criminal investigation. Sensors 9, 568–601. https://doi.org/10.3390/s90100568.

Sejnowski, T. 2020 The unreasonable effectiveness of deep learning in artificial intelligence, in: Proceedings of the National Academy of Sciences, Vol. 117, pp. 30033–30038. https://doi.org/10.1073/pnas.1907373117..

Shi, J., Yan, Q., Xu, L., Jia, J., 2016. Hierarchical image saliency detection on extended cssd. IEEE Trans. Patt. Anal. Mach. Intell. 38, 717–729. https://doi.org/10.1109/TPAMI.2015.2465960.

Silberman, N., Fergus, R., 2011. Indoor scene segmentation using a structured light sensor. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 601–608. https://doi.org/10.1109/ICCVW.2011.6130298.

Single-shot 3d shape reconstruction datasets, URL: https://figshare.com/articles/dataset/Single-shot_3D_shape_reconstruction_datasets/7636697, 2020 (accessed 21 January 2021)..

Su, X., Zhang, Q., 2010. Dynamic 3-d shape measurement method: A review. Opt. Lasers Eng. 48, 191–204. https://doi.org/10.1016/j.optlaseng.2009.03.012.

Tatarchenko, M., Richter, S.R., Ranftl, R., Li, Z., Koltun, V., Brox, T., 2019. What do single-view 3d reconstruction networks learn? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3400–3409. https://doi.org/10.1109/CVPR.2019.00352.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P., 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11, 3371–3408. https://doi.org/10.5555/1756006.1953039.

Vo, M., Wang, Z., Pan, B., Pan, T., 2012. Hyper-accurate flexible calibration technique for fringe-projection-based three-dimensional imaging. Opt. Express 20, 16926–16941. https://doi.org/10.1364/OE.20.016926.

Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., 2018. Deep learning for computer vision: A brief review. Comput. Intell. Neurosci. 13, pages. https://doi.org/10.1155/2018/7068349.

Wang, Z., Nguyen, D., Barnes, J., 2010. Some practical considerations in fringe projection profilometry. Opt. Lasers Eng. 48, 218–225. https://doi.org/10.1016/j.optlaseng.2009.06.005.

Wang, Z., Kieu, H., Nguyen, H., Le, M., 2015. Digital image correlation in experimental mechanics and image registration in computer vision: similarities, differences and complements. Opt. Lasers Eng. 65, 18–27. https://doi.org/10.1016/j.optlaseng.2014.04.002.

Wang, T., Borji, A., Zhang, L., Zhang, P., Lu, H., 2017. A stagewise refinement model for detecting salient objects in images. In: IEEE International Conference on Computer Vision (ICCV), pp. 4039–4048. https://doi.org/10.1109/ICCV.2017.433.

Wang, N., Zhang, Y., Li, Z., Fu, Y., Liu, W., Jiang, Y.-G., 2018. Pixel2mesh: Generating 3d mesh models from single rgb images. In: Computer Vision - ECCV, pp. 55–71. https://doi.org/10.1007/978-3-030-01252-6_4.

Wu, C., Wilburn, B., Matsuhita, Y., Theobalt, C. 2011. High-quality shape from multi-view stereo and shading under general illumination, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 969–976. https://doi.org/10.1109/CVPR.2011.5995388..

Xiang, Y., Mottaghi, R., Savarese, S., 2014. Beyond pascal: A benchmark for 3d object detection in the wild. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 75–82. https://doi.org/10.1109/WACV.2014.6836101.

Xie, S., Tu, Z., 2015. Holistically-nested edge detection. In: IEEE International Conference on Computer Vision (ICCV), pp. 1395–1403. https://doi.org/10.1109/ICCV.2015.164.

Yan, Q., Xu, L., Shi, J., Jia, J., 2013. Hierarchical saliency detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1155–1162. https://doi.org/10.1109/CVPR.2013.153.

Yan, S., Wu, C., Wang, L., Xu, F., An, L., Guo, K., Liu, Y., 2018. Ddrnet: Depth map denoising and refinement for consumer depth cameras using cascaded cnns. In: European Conference on Computer Vision (ECCV), pp. 155–171. https://doi.org/10.1007/978-3-030-01249-6_10.

Yan, K., Yu, Y., Huang, C., Sui, L., Qian, K., Asundi, A. 2019. Fringe pattern denoising based on deep learning, Opt. Communications 437, 148–152. https://doi.org/10.1016/j.optcom.2018.12.058..

Yang, B., Wen, H., Wang, S., Clark, R., Markham, A., Trigoni, N., 2017. 3d object reconstruction from a single depth view with adversarial learning. In: IEEE International Conference on Computer Vision Workshops (ICCVW), pp. 679–688. https://doi.org/10.1109/ICCVW.2017.86.

Yu, H., Chen, X., Zhang, Z., Zuo, C., Zhang, Y., Zheng, D., Han, J., 2020. Dynamic 3-d measurement based on fringe-to-fringe transformation using deep learning. Opt. Express 28, 9405–9418. https://doi.org/10.1364/OE.387215.

Zhang, S., 2018. High-speed 3d shape measurement with structured light methods: A review. Opt. Lasers Eng. 106, 119–131. https://doi.org/10.1016/j.optlaseng.2018.02.017.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230–6239. https://doi.org/10.1109/CVPR.2017.660.

Zheng, Y., Wang, S., Li, Q., Li, B., 2020. Fringe projection profilometry by conducting deep learning from its digital twin. Opt. Express 28, 36568–36583. https://doi.org/10.1364/OE.4104281.