Automatic depth map retrieval from digital holograms using a deep learning approach

Nabil Madali, Antonin Gilles, Patrick Gioia, Luce Morin

Automatic depth map retrieval from digital holograms using a deep learning approach

NABIL MADALI,[1,3,*] ANTONIN GILLES,[1] PATRICK GIOIA,[1,2] AND LUCE MORIN [1,3]

[1] *Institute of Research & Technology b<>com, Cesson-Sévigné, France*
[2] *Orange Labs, Rennes, France*
[3] *INSA Rennes / IETR UMR CNRS 6164 , France*
[*] *Corresponding author: nabil.madali@b-com.com*

Abstract: Information extraction from computer-generated holograms using learning-based methods is a topic that has not received much research attention. In this article, we propose and study two learning-based methods to extract the depth information from a hologram and compare their performance with that of classical depth from focus (DFF) methods. We discuss the main characteristics of a hologram and how these characteristics can affect model training. The obtained results show that it is possible to extract depth information from a hologram if the problem formulation is well-posed. The proposed methods are faster and more accurate than state-of-the-art DFF methods.

 https://doi.org/10.1364/OE.480561

1. Introduction

Holography [1] is an immersive technology that records and reproduces the emitted wavefronts of an illuminated 3D scene, providing all the depth perception cues of the human visual system [2].

The recorded hologram contains all the information that describes the acquired 3D scene. However, the extraction of 3D information directly from the hologram is not straightforward, due to the lack of spatial localization of scene objects inside the hologram plane. Recovering the scene geometry is a necessary step for several tasks, such as the segmentation and edition of the different objects contained in the scene. To directly recover the scene geometry from a hologram, researchers often use the *Depth From Focus* (DFF) [3] method on a sampled reconstruction volume. Such a volume is created from multiple reconstructions computed at different focus distances chosen in a predefined interval with a fixed depth sampling rate. Given the computed reconstruction volume, the acquisition depth can be estimated by applying focus measure operators [4] on each reconstruction and then selecting for each pixel the depth for which the focus is optimal. Numerous focus measure operators such as Entropy [5], Gray-level variance [6], Gini index [7], Tamura coefficient [8] and Wavelet transform [9] have been specifically designed for their usage on holographic reconstructions. The DFF method leads to relevant results when the focus measure operator, the reconstruction interval and the sampling rate are judiciously chosen. However, the large number of numerical reconstructions leads to a high computational cost.

With the emergence of deep learning methods, numerous research works – mainly in Holographic Microscopy where the scene is acquired at a single depth – have attempted to estimate the scene information directly from the hologram without the need for any numerical reconstruction. Given a well-designed network architecture and sufficient amount of data, research works on the subject discussed in Section 2 have shown that the trained model gives more accurate and faster results than traditional methods while maintaining to some extent their performance on new samples not seen during the training. In most cases, the trained network represents a one-to-one mapping between the hologram and a single estimated acquisition depth. Using such a network

to perform inference on complex scenes with multiple objects located at different depths – such as those used in Computer-Generated Holograms (CGH) – would require the hologram to be first segmented in an unsupervised manner into individual sub-holograms corresponding to each acquisition depth, before feeding them to the network. This is an intractable problem due to the segmentation step. In this work, we first review the possibility of learning a one-to-many mapping between the acquired hologram and a depth map of the scene. Then, a patch-based method that locally segments focused regions from a holographic reconstruction stack is presented. Finally, we compare the direct method and the patch-based in terms of generalization and scalability.

The remaining of the article is organized as follows: Section 3, first introduces the baseline approach based on the DFF method. Next, the direct mapping between the hologram and the scene depth map is introduced. Finally, the proposed patch-based approach and the used network architectures are detailed. In Section 4, a series of experiments are conducted to validate the proposed approach, then its advantages and limitations are discussed in Section 5.

2. Related work

Depth estimation is a well-studied problem in *Digital Holographic Microscopy* (DHM). To recover the acquisition depth value from a given hologram, most methods rely on a set of holographic reconstructions computed at different depths from the hologram plane and then estimate a sharpness value for each reconstruction using a Focus Measure (FM) operator [4]. The maximum of the sharpness curve indicates the distance at which the microscopic object (cell, molecule, etc.) is located from the hologram plane. Several focus operators designed for DHM have been proposed in the literature; a fairly exhaustive review can be found in [10].

With the advent of deep learning, several research works tried to solve this problem directly, by designing end-to-end architectures that directly use the hologram without requiring additional pre-processing, making it possible to obtain more precise and faster results than traditional methods.

One of the pioneer works on the subject was done by Pitkaaho *et al.* [11]. It uses the AlexNet [12] network architecture to estimate the in-focus depth from a set of holographic reconstructions computed at manually selected axial positions. Each recorded hologram is manually preprocessed to remove zero and twin order terms. Then, twenty-one holographic reconstructions are computed with a fixed reconstruction step of 10mm within a $\pm$100mm range around the ground truth in-focus depth. The reconstruction amplitudes are resized, randomly cropped, and augmented using different rotation angles, before feeding them into the network. The network is supervised to predict where the focus is optimal among the twenty-one reconstructions. Experimental results show that the network generalizes well using different cell types when the hologram is recorded using the same optical setup as for the training. However, it fails to generalize under different magnification and lighting conditions during hologram acquisition.

Subsequently, Ren *et al.* [13] proposed to formalize the depth estimation as a classification problem. Given the input hologram, the CNN network is supervised to predict a label which represents the discretized object distance from the hologram plane. The network is trained on a large-scale dataset, composed of 1000 holograms acquired optically from the local area of the negative USAF 1951 at different lateral and axial positions. Each hologram was recorded at one of five possible recording distances, acting as the targeted labels. The experimental results show that through the training, the network learns the relevant characteristics to cluster each input hologram into one of the five possible classes, outperforming the results obtained using kNN and SVM by a large margin for both validation and test, with almost 98% accuracy.

The same authors extended their work in [14] by reformulating the depth estimation as a regression problem. The used CNN architecture consists of five convolution blocks, each block containing a convolution layer followed by a batch normalization and maximum pooling layers. The output of the last convolution block is vectorized and passed through two fully connected

layers to produce the estimated focus depth. The whole network is supervised using an ℓ_2 norm between the estimated and ground truth values. The authors reviewed the network performance on amplitude and phase objects using holograms acquired optically from local areas of USAF 1951 and several biological specimens. For amplitude objects, 500 holograms with ten lateral positions were used, and for phase objects, 2000 holograms were used with five different lateral positions and four magnification levels. The experimental results showed that CNN outperforms MLP, k-nn, and SVM models for both amplitude and phase objects. The network shows a reasonable ability to generalize to different optical recording exposures and axial distances inside the training range. However, outside the training range, the network fails to produce the correct focal distance estimate. Overall, the proposed network is faster, more accurate, and with fewer hyper-parameters than conventional auto-focusing methods.

Shimobaba *et al*. [15] improved the previous work by introducing the hologram power spectrum as an additional input data to the CNN, arguing that it gives more relevant information than the raw interference pattern for predicting focused depth. To verify this assumption the authors trained two identical networks, one for each input type, with holograms computed from Caltech-256 datasets [16]. The obtained results demonstrate that the network trained on the raw interference model pattern is unable to give a correct prediction on the validation set. On the other hand, the network trained on the power spectrum not only performs better on the training set but also maintains its performance on the validation set. The proposed approach yields faster and better results than Tamura coefficient [17] method without the need for a computationally expensive deep search.

In the previously cited works, the network is supervised to predict a single acquisition depth per hologram. Given a pre-trained model, the inference for a complex scene containing multiple objects requires the hologram to be segmented in an unsupervised manner into a set of sub-holograms, each corresponding to a unique acquisition depth. Then, the inference results of the different sub-holograms must be merged to obtain the depth map of the scene. This is an intractable problem due to the lack of spatial localization in the hologram plane.

In the following section, we describe our deep learning-based methodology to estimate the depth map from a computer-generated hologram of a complex scene with multiple focal distances.

3. Methodology

In this section, we first recall the DFF method and how to use it to extract multiple depths per hologram. Then, we propose two different deep learning-based approaches for retrieving the scene geometry from a given hologram, illustrated in Figures 2 and 3.

3.1. Depth From Focus (DFF) Principle

The following section briefly describes the DFF approach which is often applied in holographic microscopy [10] where a single depth value is expected per hologram. The DFF approach will serve as a baseline model to evaluate the performances of the proposed approaches.

3.1.1. Single depth extraction per hologram

Given a hologram denoted by $H \in \mathbb{C}^{L\times L}$, a reconstruction volume is acquired by computing a set of numerical reconstructions on planes that are parallel to the hologram and located in a predefined depth interval $[z_{\min}, z_{\max}]$, where $z_{\min}$ and $z_{\max}$ are the minimal and maximal depths of the scene respectively. Each numerical reconstruction is computed by the Angular Spectrum Method [2] defined as

$$\mathcal{P}_{z_i}\{H\} = \mathcal{F}^{-1}\left\{\mathcal{F}(H)e^{j2\pi z_i\sqrt{\lambda^{-2}-f_x^2-f_y^2}}\right\}, \tag{1}$$

where f_x and f_y are the spatial frequencies along the X and Y axis, λ is the acquisition wavelength, and z_i is the reconstruction depth, given by

$$z_i = \frac{z_{\max} - z_{\min}}{N} i + z_{\min}. \tag{2}$$

The set of numerical reconstruction amplitudes constitutes the reconstruction volume defined as

$$\Omega = \{\left|\mathcal{P}_{z_0}\{H\}\right|, \left|\mathcal{P}_{z_1}\{H\}\right|, ..., \left|\mathcal{P}_{z_N}\{H\}\right|\}. \tag{3}$$

When the original scene is located at a single depth, this depth value can be recovered by evaluating the depth for which the reconstruction is sharpest with a focus measure operator FM according to

$$d = \underset{i\in[1,N]}{\arg\max} \left\{FM(\left|\mathcal{P}_{z_i}\{H\}\right|)\right\}. \tag{4}$$

3.1.2. Extraction of several depths per hologram

For a complex scene with multiple objects located at different depths, each reconstruction plane in the reconstruction volume must first be decomposed [18] into either non-overlapping or overlapping patches of predefined resolution ($s \times s$). An optimal decomposition ensures that only one depth value is expected per patch $R_{m,n,i}$ given by

$$R_{m,n,i}(u,v) = \left|\mathcal{P}_{z_i}\{H\}\right|(m+u-s/2, n+v-s/2). \tag{5}$$

The depth $d_{m,n}$ is therefore estimated independently for each set of patches $\{R_{m,n,i} \mid i \in [1,N]\}$ as

$$d_{m,n} = \underset{i\in[1,N]}{\arg\max} \left\{FM(R_{m,n,i})\right\}, \tag{6}$$

yielding a depth map containing for each pixel (m,n) the depth at which the focus of the corresponding patch centered around the pixel is optimal.

The numerical error between the estimated and ground truth depths depends on the chosen reconstruction interval, sampling rate, and focus measure operator. The reconstruction interval and sampling rate control the precision of the committed numerical error. With a reconstruction interval closer to the optimal focal plane and a finer sampling rate, the numerical error is small, since the reconstructions are performed closer to the optimal solution. On the other hand, a larger reconstruction interval or low sampling rate induces a high numerical error, since most reconstructions are far from the optimal focus plane.

Besides the holographic reconstruction parameters, the major drawbacks of the DFF approach are the optimal decomposition of the reconstruction volume and the choice of the patch size. Figure 1, illustrates two cases that can be encountered in practice. In the first case, the extracted patch is made up of four planar sub-regions, each with a different acquisition depth. If this patch is used to estimate the depth of the central pixel, there is a $1/4$ chance that the estimate will be accurate; however, this probability can be raised by reducing the patch size. The patch defined by the red dashes is the most optimal and ensures a single depth value per patch. In the second scenario, the patch is divided into several sub-regions, each with a distinct geometry. It is therefore not possible to accurately predict the depth of the central pixel by reducing the size of the patch because it will always contain multiple sub-regions, creating uncertainty in the depth prediction.

To alleviate these drawbacks, alternative methods based on deep learning are described in the next section.

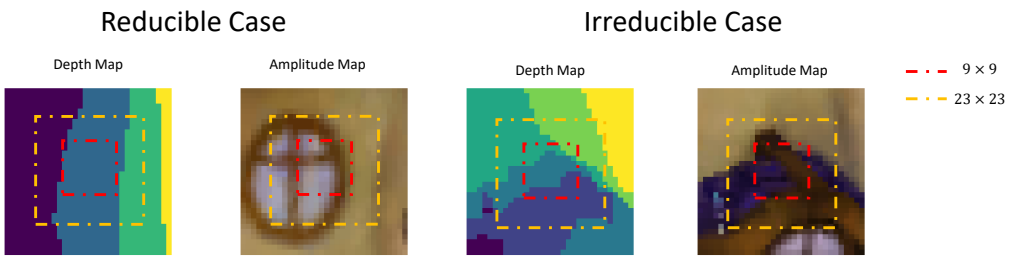

Fig. 1. Illustration of the main drawbacks of the DFF approach using two cropped patches of size (32×32) extracted at two different positions. The dotted lines correspond to the new spatial boundaries of the cropped patches with a crop size of (23×23) and (9×9) respectively. In the first case, the depth value of the central pixel can be correctly estimated if the patch size is considerably reduced. In the second case, regardless of the chosen patch size, the depth prediction is subject to uncertainty.

3.2. Proposed approaches based on CNN

In this section, two learning-based approaches are discussed. First, a direct approach is illustrated in Figure 2 that directly uses the hologram as input and is supervised to predict the scene depth map. Then, a patch-based method illustrated in Figure 3 alleviates the main drawback of the DFF approach.

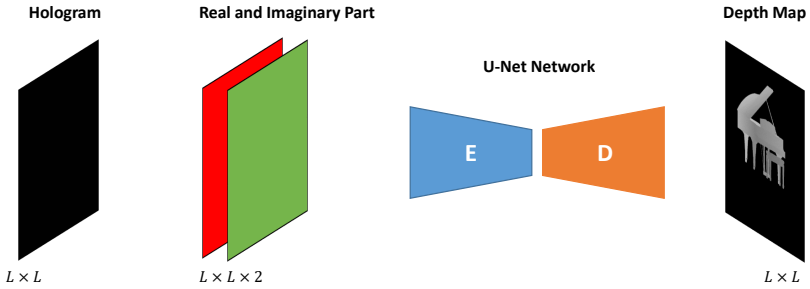

Fig. 2. Illustrations of the direct mapping approach where the hologram is directly used as the input of the network.

3.2.1. Direct mapping approach from the hologram to the depth map

The first learning-based approach that was considered, is a direct mapping denoted by $\mathcal{G}_1$ from the recorded hologram H to the focused depth $d \in \mathbb{R}^{L \times L}$, such that

$$d = \mathcal{G}_1\{H\}, \qquad H \in \mathbb{C}^{L \times L} \tag{7}$$

which is a reasonable assumption since the hologram contains both the intensity and depth information of the acquired 3D scene.

In previous works, $\mathcal{G}_1$ is a one-to-one mapping between the recorded hologram H and a single acquisition depth d. In the present work, the scenes are composed of several objects having a different shape and location in 3D space. Therefore, during the training process, the network must first extract from H the relevant features to segment the scene parts that are at the same depth and then predict their correct focus distance. Learning $\mathcal{G}_1$ from a scene comprising several objects located at different positions requires a much larger dataset, by varying not only the position but also the shapes of the objects that compose the scene. The slightest change in scene

geometry or aberration on the hologram can prevent the network from predicting relevant results. It is therefore necessary to ensure that all input parameters are covered during training, thus guaranteeing relevant results in the testing phase.

Besides the need for a very large number of training samples, a direct mapping approach lacks scalability. Indeed, the computational complexity for network training and inference increases linearly with the input resolution.

In order to design an efficient and scalable direct approach method, it is necessary to maintain a fixed spatial resolution for the network input. Therefore, for a high-resolution input, the network training and inference should be performed locally on different sub-parts of the initial input and then merged together to obtain the final results. Training the network locally greatly reduces the computational cost and GPU usage, and allows for a greater generalization ability, because the structural changes that can be observed in a local area are limited and can therefore easily be learned. However, since there is no spatial localization in the hologram plane, it is not possible to segment the hologram into independent sub-parts for network training and inference. Indeed, each pixel of the targeted depth map depends on every pixel of the hologram.

For this purpose, a second learning-based method that combines the benefits of the classic DFF method and the direct approach while alleviating their drawbacks is presented in the next section.

3.2.2. Patch-based approach

As stated previously, to design an efficient learning-based approach, it is important to perform the network training and inference locally using a fixed and low input/output resolution, allowing a faster and more scalable model with good generalization abilities. To design such an approach, the hologram cannot be directly segmented due to the lack of spatial localization of the scene objects in the hologram plane. Therefore, the hologram is first converted into a reconstruction volume to retrieve the spatial locations of the scene objects in the reconstruction images, then a mapping is learned between the cropped reconstruction volumes and the corresponding targeted depth map as illustrated in Figure 3.

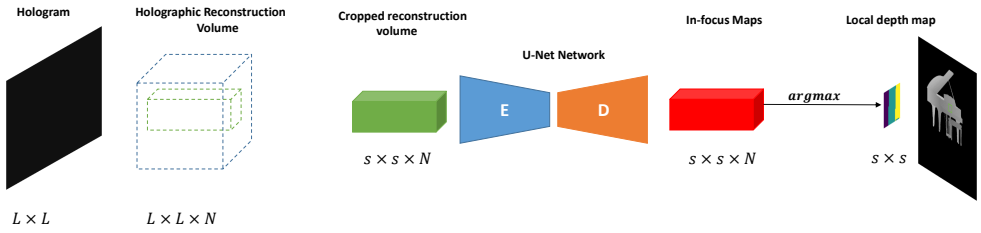

Fig. 3. Illustrations of the patch-based approach where a reconstruction volume is first constructed from the hologram, then the network is fed with a cropped reconstruction volume to predict the regional depth map.

The proposed method has some similarities with the DFF method. First, for both approaches the same reconstruction volume Ω given in Eq. 3 is used. Secondly, as for the DFF method, the goal of the neural network is to estimate for each pixel the depth for which the focus is locally optimal. The major difference between the two approaches is the used FM function in Eq. 6, as detailed in the following.

In Eq. 6 the FM function can only evaluate the focus level of the given patch and does not provide which part of the patch is in focus. Therefore, in order to compute a complete depth map, the inference operation must be performed locally for each pixel of the reconstruction volume,

using a patch centered around each pixel with an optimal patch size to guarantee a single depth per patch. Lets consider the case where the FM function is replaced with a learned network denoted by $\mathcal{G}_2$. In that case, the network does not only evaluate the level of focus but also segments the in-focus and out-of-focus regions of the given patch, as illustrated in Figure 4. Consequently, the inference operation can be performed using disjoint patches, and the targeted depth map can be inferred from the reconstruction distance used for each segmented region. Thus, the number of operations required to compute a complete depth map is reduced from $L \times L \times N$ operations to $\frac{L}{s} \times \frac{L}{s} \times N$ operations plus additional gains from GPU parallelization.

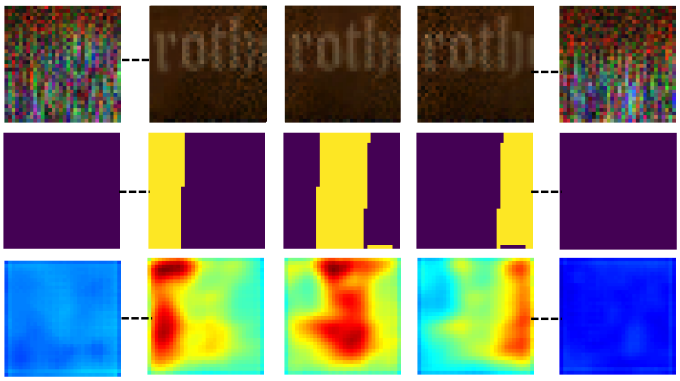

Fig. 4. First row: input reconstruction volume $\{\Lambda_{p,q,i} \mid i \in [1, N]\}$. Second row: ground truth focus maps. Third row: predicted focus maps $\{\hat{I}_i \mid i \in [1, N]\}$.

The decomposition into a set of non-overlapping patches can be formulated as:

$$\Theta = \left\{ \Lambda_{p,q,i} \quad \middle| \quad p \in \left[1, \left\lfloor \frac{W}{s} \right\rfloor\right], \quad q \in \left[1, \left\lfloor \frac{H}{s} \right\rfloor\right] \right\}, \tag{8}$$

$$\Lambda_{p,q,i} = \left|\mathcal{P}_{z_i}\{H\}\right|(ps+u, qs+v), \quad u \in [0, s-1], \quad v \in [0, s-1], \quad i \in [1, N] \tag{9}$$

where $\lfloor a \rfloor$ is the greater integer value smaller than or equal to a, H and W are respectively the height and width of the reconstruction volume, and s in the patch size.

Given this new decomposition, Eq. 6 can be reformulated as follows:

$$\hat{I}_i = \mathcal{G}_2\{\Lambda_{p,q,i}\}, \tag{10}$$

$$\hat{d}_{p,q} = \sum_{i=1}^{N} i \times [\hat{I}_i] \tag{11}$$

where $\hat{I}_i$ is the predicted in-focus map corresponding to the holographic reconstruction performed at the distance z_i, $[\hat{I}_i]$ is the rounded version of $\hat{I}_i$, and $\hat{d}_{p,q}$ is the predicted depth map.

In order to train the network $\mathcal{G}_2$, two problems arise. First, it is necessary to sample a large number of patches with a positive sample that includes focused regions with different textures, positions and shapes, as well as negative samples without focused regions. Secondly, to guarantee the occlusion condition, only one maximum focus value can be allowed per pixel. If the different images that compose the cropped volume $\{\Lambda_{p,q,i} \mid i \in [1, N]\}$ are used independently, this can lead to several extreme values per pixel and therefore to uncertainty when choosing the optimal depth value.

To alleviate these problems, all the images of the cropped reconstruction volume are fed to the network as unique batch, and the network is supervised using cross-entropy loss to predict a single in-focus depth per pixel:

$$\mathcal{L} = \sum_{u,v} \sum_{i=1}^{N} \mathbb{1}_{\{d(u,v)=z_i\}} \cdot log\hat{I}_i(u,v) \tag{12}$$

After the network training is complete, the predicted depth per pixel can be extracted according to

$$\hat{d}_{p,q}(u,v) = \underset{i\in[1,N]}{\arg\max}\, \hat{I}_i(u,v) \tag{13}$$

and the associated color intensity value as,

$$\hat{Im}_{p,q}(u,v) = \Lambda_{p,q,\hat{d}_{p,q}(u,v)}(u,v) \tag{14}$$

In cases of occlusion where the focus has multiple peaks, using the k-th maximum value in Equation 13 may not be reliable. This is because the network is designed to identify a single optimal focus plane for each input reconstruction volume, which results in a focus curve with a single peak depth value. To address this issue, the network can be applied to reconstruction volumes computed using a sliding window over the reconstruction interval. This approach allows for the extraction of peak values incrementally, rather than relying on a single peak value from the entire reconstruction volume. It is important to ensure that the extracted peaks pass a certain threshold to ensure reliable results.

The obtained RGB-D image $(\hat{Im}, \hat{d})$ provides a comprehensive representation of the scene geometry through the use of holographic reconstructions. This is useful information for many applications, such as holographic video compression where the patch-based method can be used on consecutive holographic video frames to extract their RGB-D representation. Bi-dimensional motion vectors on the XY axis can be estimated using optical flow techniques on the RGB data, and additional motion along the Z axis can be estimated using the depth information. These motion vectors can be incorporated into a compression framework to improve the efficiency and quality of the encoded holographic video data.

Requiring the full cropped volume for efficient network training may limit the used patch size and the number of reconstructions. It is, therefore, necessary to have a trade-off between the used patch size and the number of holographic reconstructions in order to avoid high GPU usage. In this new formulation, the problem is well-posed due to the out-of-focus behavior of CGH. When modifying the reconstruction distance on CGH, only the parts with a depth at the reconstruction distance are sharp while the remaining regions are contaminated by speckle noise as shown in Figure 5. Therefore, the detection of focused regions is similar to the removal of the noisiest regions.

3.3. Network Architecture

The functions $\mathcal{G}_{1,2}$ are implemented as CNN network, based on the U-Net architecture [19] and is shown in Figure 6. The network is composed of two sub-networks, a contracting path and an expansive path. The network can be mathematically described as

$$X^{i,j} = \begin{cases} \mathcal{D}(\mathcal{H}(X^{i-1,j})), & j = 0 \\ \mathcal{H}(\left[X^{i,0}, \mathcal{U}(X^{i+1,j-1})\right]), & j = 1 \end{cases}, \tag{15}$$

where $X^{i,j}$ is the initial network input, $\mathcal{H}$ is a convolution block shown on the right side of Figure 6, $\mathcal{D}$ and $\mathcal{U}$ are 2×2 Max-pooling and bilinear interpolation operations, and [] is a

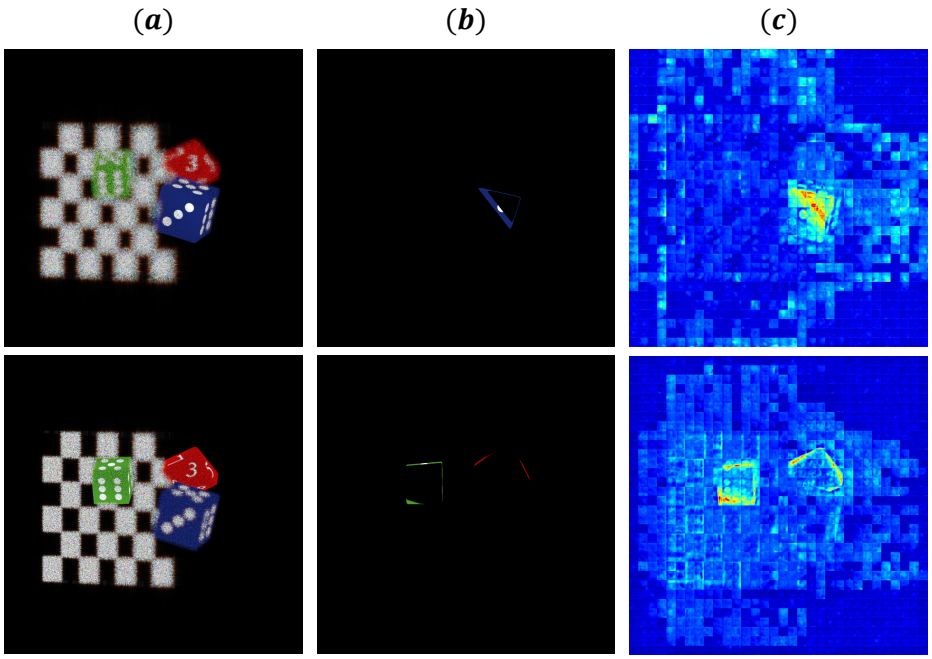

Fig. 5. Illustration of the predicted in-focus regions on validation sample. (a) Amplitude of holographic reconstruction at a given distance. (b) The ground truth in-focus regions. (c) The predicted in-focus scores.

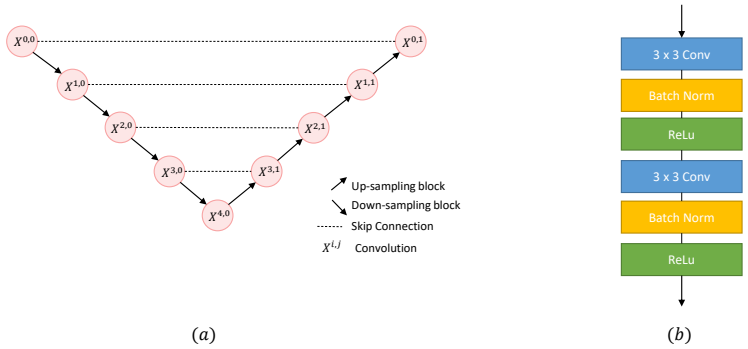

Fig. 6. Illustration of the network architecture in (a) and the convolution block $\mathcal{H}$ in (b)

concatenation operator. The depth resolution for the contraction path is [32,64,128,256,512] and [256,128,64,32,1] for the expansive path. The segmentation network shares the same architecture but uses half the depth resolution in order to avoid over-fitting.

The skip-link connections play an important role, especially on holograms that contain a lot of redundant information, since each light wave scattered by each scene point contributes to every pixels during hologram recording. With the skip-link connection, at each stage of the decoder, the network can use both the local and global features extracted from the hologram, therefore, the depth value at each pixel is learnt using a greater amount of information than using a standard auto-encoder.

4. Experimental results

4.1. Experimental Setup

To train the neural networks, a large-scale dataset composed of 4200 RGB holograms obtained from three different scenes (Piano, Table, Woods shown in Figure 10) has been acquired using a layer-based method [20], with a resolution of 1024×1024, a pixel pitch of $6\mu m$ and 1400 acquisition angles. For the evaluation phase, two sets are used: a test set composed of 300 holograms computed using 100 additional acquisition angles for each training scene (Piano, Table, Woods), and a validation set composed of 200 holograms recorded from two unseen scenes (Cars, Dices) during the training phase.

The two networks for the direct mapping and patch-based approaches were trained and tested on the same training and validation sets. For direct mapping, the real and imaginary parts of the RGB channels are stacked together and fed directly to the network. The network training is supervised using the ℓ_1 norm with an additional spatial gradient loss to alleviate the blurred effect produced by the CNN. More formally

$$\mathcal{L} = \frac{1}{N}\left[\sum_{i=1}^{N} \left\|d - \hat{d}\right\|_1 + \sum_{i=1}^{N} (\nabla_x \left\|d - \hat{d}\right\|_1 + \nabla_y \left\|d - \hat{d}\right\|_1)\right], \tag{16}$$

where N_x, N_y are the height and width of the depth map, and d and $\hat{d}$ are the ground truth and estimated depth maps respectively.

In the patch-based approach, a set of holographic reconstructions have been computed at unique depths of the ground truth depth map for each hologram. The resulting reconstruction volume is decomposed spatially into non-overlapping patches. All the patches extracted at the same spatial position are fed into the network, which is supervised using the cross-entropy loss to predict the correct in-focus maps.

The two networks have been trained for 200 epochs using the Adam optimizer and learning decay with a gamma factor of 0.8 every 10 epochs and a starting learning rate of 0.1. The networks are compared to nine commonly used focus operators listed in Table 1. The operators are applied on windows of size 17×17 which gave the best results on previous experiments.

Focus operator	Abbr.	Focus operator	Abbr.
Energy of Laplacian [21]	LAPE	Modified Laplacian [22]	LAPM
Variance of Laplacian [23]	LAPV	Diagonal Laplacian [24]	LAPD
Variance of wavelet coefficients [25]	WAVV	Ratio of the wavelet coefficients [25]	WAVR
Graylevel variance [26]	GLVA	Normalized Graylevel variance [27]	GLVN
Image contrast [28]	CONT		

Table 1. Abbreviations of focus measure operators used in experiments.

4.2. Results

Table 2 reports the obtained results on the validation and test sets. First, we observe that for learning-based approaches, only the patch-based model maintains its performance on both the validation set (Piano, Table, Cars) and test set (Dices, Cars). The direct mapping approach results in close performance between the training set and the validation set but fails to generalize on the test set. The lack of generalization can be explained by the problem formulation which is not localized. During training, the network learns from a set of training data that covers only a part of all possible scene shapes and positions. Therefore, for test scenes with sizes, shapes, and spatial positions unseen during training, the direct mapping network fails to give relevant results

	Piano	Table	Woods	Dices	Cars
Direct approach					
U-Net	**6.94**	**8.48**	8.62	19.93	23.92
Patch-based approach					
U-Net	**0.66**	**1.12**	**0.73**	**1.21**	**3.48**
Laplacian-based operators					
LAPE	11.58	25.81	16.35	**5.83**	32.56
LAPM	18.89	16.48	23.158	27.0	51.26
LAPV	**4.12**	**9.28**	**3.82**	**2.84**	**17.07**
LAPD	23.88	19.93	26.81	26.81	54.16
Wavelet-based operators					
WAVV	9.95	11.24	**7.19**	13.26	23.88
WAVR	8.16	8.98	7.45	8.76	**14.44**
Miscellaneous operators					
GLVN	17.90	15.10	20.70	25.75	51.65
GLVA	21.40	17.10	19.38	24.12	51.62
CONT	16.33	14.89	19.51	21.30	47.13

Table 2. L1 error using different approaches for validation and test sets.

and tends to smooth the final results given an average estimated value of all pixels belonging to the scene.

The patch-based approach maintains its performance on the test set, due to the fact that the different patterns that are observed during the test phase are similar to those observed during the training phase. In the case where the observed patterns differ greatly from the training phase, the problem being well-posed thanks to the property of the CGH, the network gives fairly consistent results.

For the DFF approach, there is no operator that maintains its performances for the different scenes; the performance of the operators will indeed depend on the scene geometry. For scenes with occlusions, some objects will be positioned one after the other, resulting in several levels of focus on the patches extracted at the objects' borders; in this case, predicting a single depth for the central pixel results in poor performance. The ratio between the textured and non-textured patches also plays a role on the final performance, some operators will perform better on highly textured areas but will fail on flat areas without texture. Overall, the Laplacian variance and Wavelet-based approaches are the best-performing operators.

Figures 7, 8, 10, 11 give some visual results. From these figures, we observe that the obtained results with the direct mapping are smooth and fail to deal with the occlusions correctly. Indeed, when two objects are one behind the other, their borders are smooth, resulting in the integration of the first object into the second. In addition, while the LAPV operator yields a reasonable performance, it fails in regions with high structural variability. When a patch is extracted over a region where the depth changes rapidly, the patch has a high probability of containing multiple levels of focus, so predicting a single depth for the entire patch will lead to poor performance. The patch-based approach is well suited to this type of scenario. Segmenting locally for each holographic reconstruction only the parts that are in focus, allows multi-level focus processing within the same patch, while maintaining consistent edges and contours between the different extracted regions.

In summary, the proposed patch-based approach is more efficient than the classic DFF method

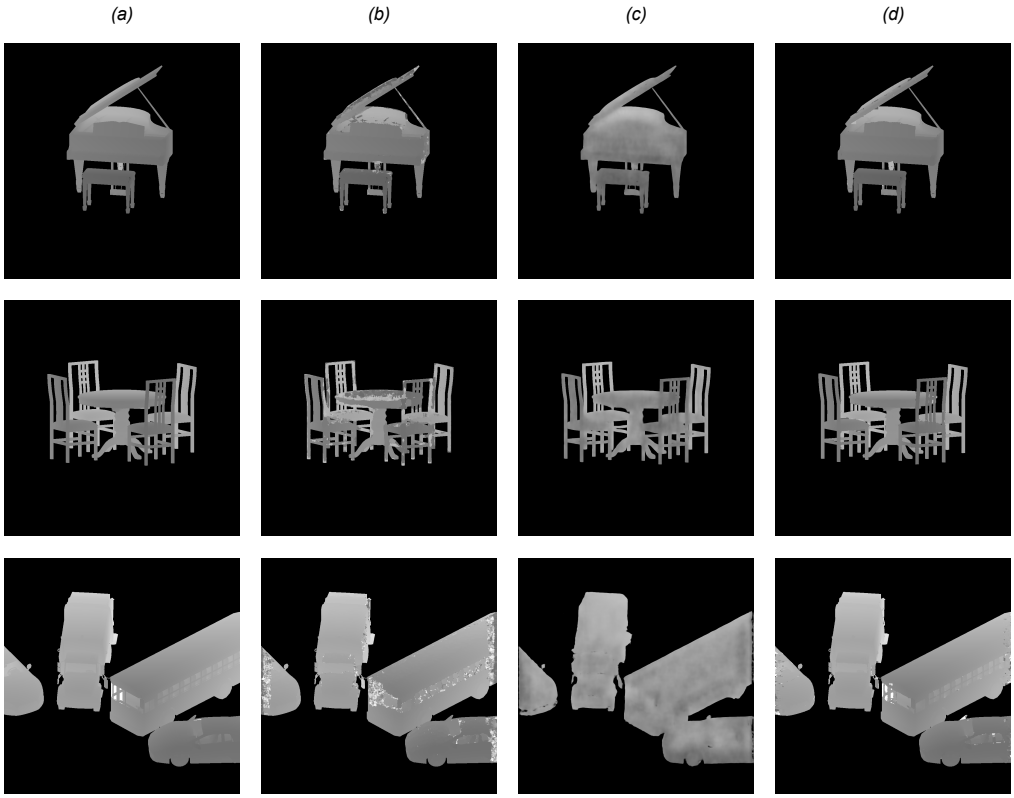

Fig. 7. Each row contains a sample taken from our datasets. (a) The ground truth all in-focus image. (b) The depth map estimated using the LAPV operator. (c) The depth map estimated using direct mapping. (d) The depth map estimated using the proposed approach.

as it can use a single batch inference to determine the depth map for each crop of the reconstruction volume. Additionally, it is more accurate in segmenting the different sub-regions of the cropped reconstruction volume, and does not require further post-processing or adaptive per-pixel patch sizes to maintain depth continuity in the final depth map. The network is trained to identify the relevant characteristics of an in-focus image using fixed-size input patches and properly segment it. Furthermore, the network ensures that the focus level of each pixel follows a perfect Gaussian distribution, with peaks corresponding to the optimal focus, and is therefore not subject to polarity changes when the patch size or reconstruction range is too large.

The direct CNN approach has a faster computation time (5.64 seconds per hologram), but it lacks the ability to generalize well compared to the patch-based approach (1.31 seconds per holographic reconstruction and 26.05 seconds for processing a reconstruction volume composed of 256 holographic reconstruction). Both approaches were trained and tested using the same dataset, so this difference in generalization ability cannot be attributed to differences in the quality or quantity of the data. Instead, it is likely due to the different approaches to feature extraction and problem formulation. The direct CNN approach is ill-posed, which causes the network to overfit the data and extract features that are specific to the training set, rather than general features that can be applied to new, unseen data. As a result, the network performs well on the training set but struggles to make accurate predictions on new data because it has not learned to generalize effectively. In contrast, the patch-based approach is able to extract more generalizable features,

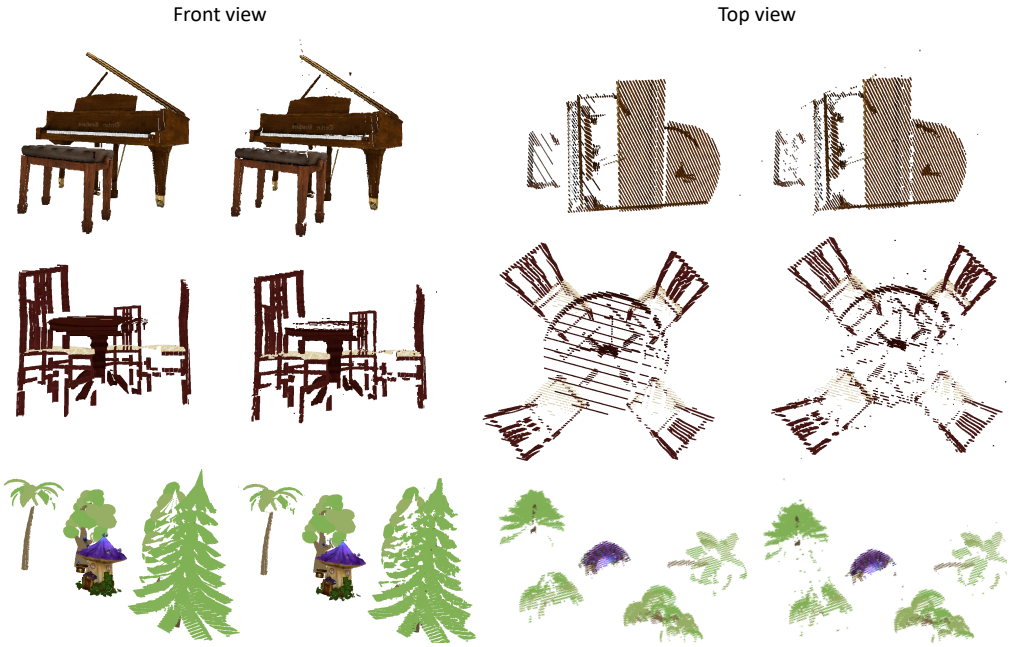

Fig. 8. Illustration of pointclouds predicted using the patch-based method on samples from the test set. On the right the ground truth points cloud and on the left the prediction using the proposed approach.

resulting in better generalization ability and the ability to make accurate predictions on new data.

In addition to generated data, the patch-based approach was applied to some of the IRT-bcom dataset [29], which was generated using alternative hologram generation techniques (as shown in the Figure 9). The obtained results indicate that the network was able to accurately extract the correct depth value, despite some parts of the depth map being poorly estimated. This is due to the smaller number of reconstruction distances samples leading to the sampled reconstruction distances being too far from the optimal focus planes.

5. Conclusion

In this paper, we propose two different learning-based approaches for depth estimation from CGH and compare their performances to classical DFF methods. In the first direct mapping approach, the network takes the hologram as input and it is supervised to predict the correct scene depth map. In the second more localized approach, the network takes as input a cropped reconstruction volume and is supervised to predict a local depth map.

Overall, the patch-based approach gives the most accurate depth prediction results. It enables to segment only the parts that are in focus on each holographic reconstruction, allowing multi-level focus processing within the same patch, while maintaining consistent edges and contours between the different extracted regions.

Although the proposed approach is more efficient than the classic DFF method and the direct mapping approach, it has some limitations. Firstly, the GPU consumption increases linearly with the number of holographic reconstructions and the patch size. As the accuracy of the estimated depth is directly related to the number of holographic reconstructions, a trade-off between accuracy and computational complexity must be considered. Similarly, the robustness of the depth map is directly related to the patch size: a larger patch size will provide a more

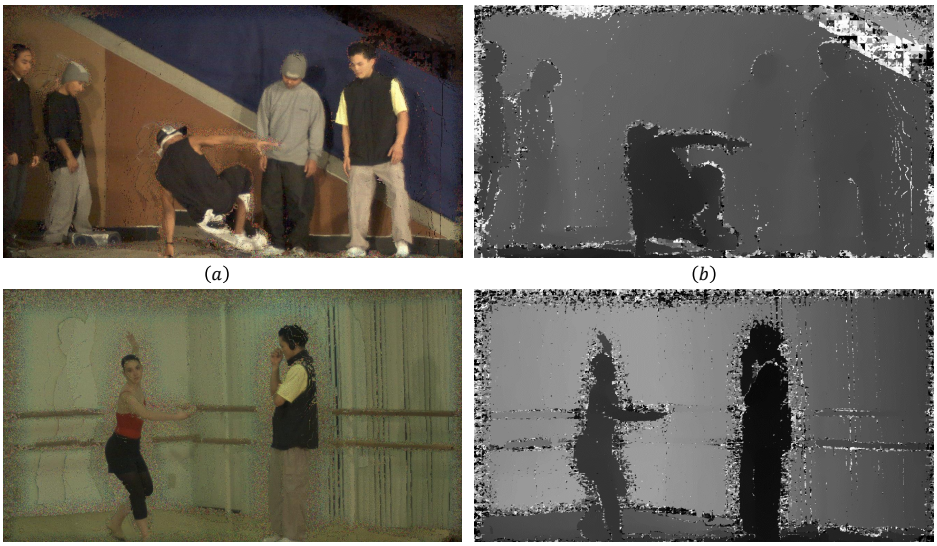

Fig. 9. Depth ((b) and (d)) and focus maps ((a) and (c)) extracted using the patch-based method on samples (Breakdancers1080p and Ballet1080p) from the IRT-bcom dataset. The network was used on a holographic reconstruction volume consisting of 256 holographic reconstructions performed at the uniformly sampled distance between the maximum and minimum depth values provided in each hologram metadata. The results indicated that the network was able to accurately extract depth information, but some depth map regions were estimated poorly. These areas could potentially be improved by implementing more precise holographic distance sampling.

accurate estimation, but at the cost of increased computational complexity. Secondly, the network performance is weaker along the edges of the patches, which can lead to discontinuities in the final depth map. Additionally, like DFF methods, the resulting depth map requires the calculation of a binary mask to distinguish between foreground and background objects, which adds additional computational costs.

Future research will focus on investigating novel learning-based techniques that directly learn a transformation in an intermediate space from the hologram where the scene objects are spatially localized and then, in a subsequent step, predict their depths.

Funding. Agence Nationale de la Recherche (ANR-A0-AIRT-07).

Disclosures. The authors declare no conflicts of interest.

Data availability. No data were generated or analyzed in the presented research.

References

1. D. Gabor, "A new microscopic principle," Nature **161**, 777–778 (1948).
2. J. W. Goodman, "Introduction to fourier optics," Introd. to Fourier optics, 3rd ed., by JW Goodman. Englewood, CO: Roberts & Co. Publ. 2005 **1** (2005).
3. P. Grossmann, "Depth from focus," Pattern Recogn. Lett. **5**, 63–69 (1987).
4. R. A. Muller and A. Buffington, "Real-time correction of atmospherically degraded telescope images through image sharpening," J. Opt. Soc. Am. **64**, 1200–1210 (1974).
5. J. Gillespie and R. A. King, "The use of self-entropy as a focus measure in digital holography," Pattern Recognit. Lett. **9**, 19–25 (1989).
6. L. Ma, H. Wang, Y. Li, and H. Jin, "Numerical reconstruction of digital holograms for three-dimensional shape measurement," J. Opt. **6**, 396–400 (2004).

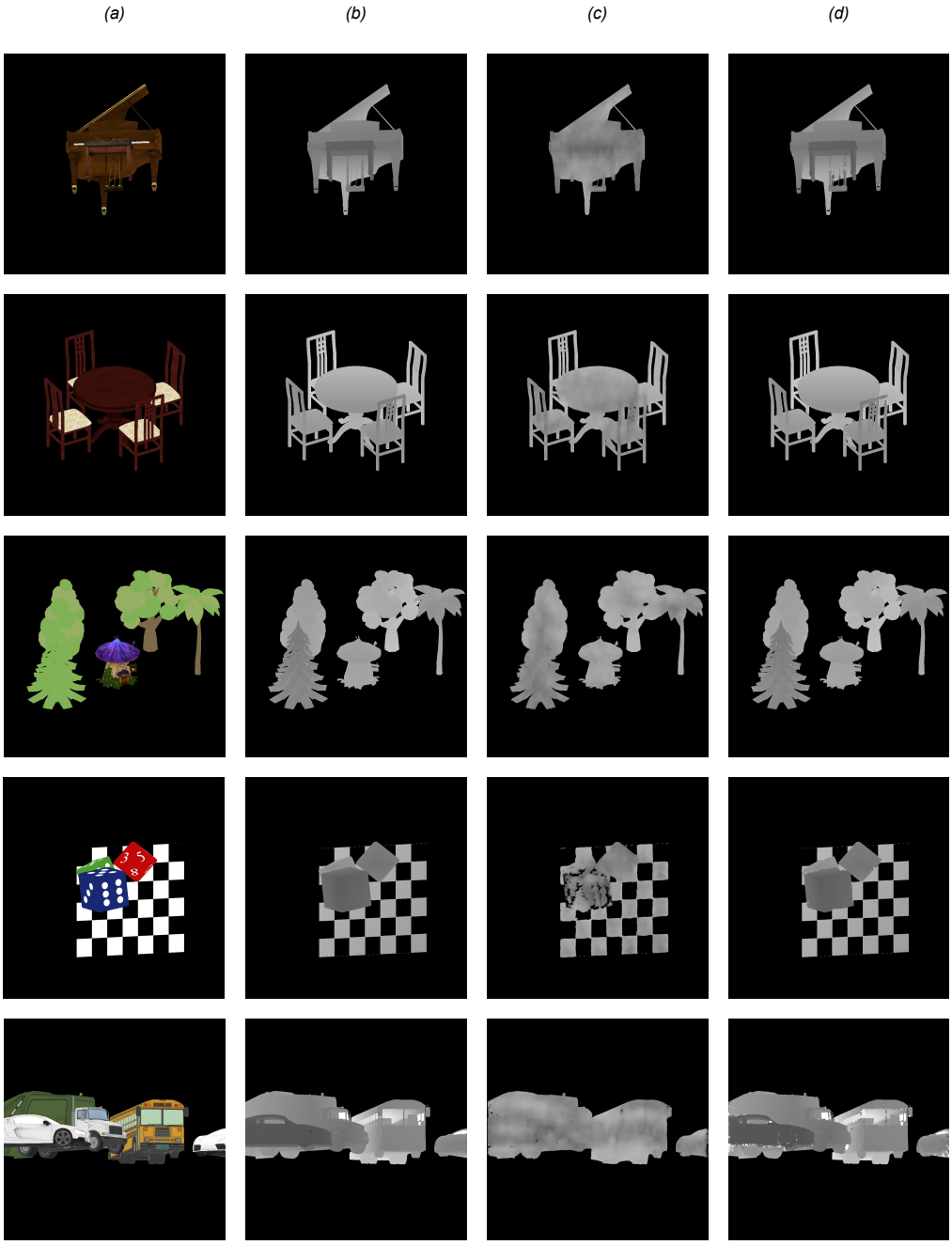

Fig. 10. Each row contains a sample taken from our datasets. (a) The ground truth all in-focus image. (b) The ground truth depth map. (c) The depth map estimated using direct mapping. (d) The depth map estimated using the proposed approach.

Fig. 11. On the right the ground truth points cloud and on the left the prediction. On the right the ground truth points cloud and on the left the prediction using the proposed approach.

7. D. Zonoobi, A. A. Kassim, and Y. V. Venkatesh, "Gini index as sparsity measure for signal reconstruction from compressive samples," IEEE J. Sel. Top. Signal Process. **5**, 927–932 (2011).
8. P. Memmolo, C. Distante, M. Paturzo, A. Finizio, P. Ferraro, and B. Javidi, "Automatic focusing in digital holography and its application to stretched holograms." Opt. letters **36 10**, 1945–7 (2011).
9. M. Liebling and M. A. Unser, "Autofocus for digital fresnel holograms by use of a fresnelet-sparsity criterion." J. Opt. Soc. Am. A, Opt. image science, vision **21 12**, 2424–30 (2004).
10. J. T. Sheridan, R. K. Kostuk, A. F. Gil, Y. Wang, W. Lu, H. Zhong, Y. Tomita, C. Neipp, J. Francés, S. Gallego, I. Pascual, V. Marinova, S.-H. Lin, K.-Y. Hsu, F. Bruder, S. Hansen, C. Manecke, R. Meisenheimer, C. Rewitz, T. Rölle, S. Odinokov, O. Matoba, M. Kumar, X. Quan, Y. Awatsuji, P. W. Wachulak, A. V. Gorelaya, A. A. Sevryugin, E. V. Shalymov, V. Y. Venediktov, R. Chmelik, M. A. Ferrara, G. Coppola, A. Márquez, A. Beléndez, W. Yang, R. Yuste, A. Bianco, A. Zanutta, C. Falldorf, J. J. Healy, X. Fan, B. M. Hennelly, I. Zhurminsky, M. Schnieper, R. Ferrini, S. Fricke, G. Situ, H. Wang, A. S. Abdurashitov, V. V. Tuchin, N. V. Petrov, T. Nomura, D. R. Morim, and K. Saravanamuttu, "Roadmap on holography," J. Opt. **22**, 123002 (2020).
11. T. Pitkäaho, A. Manninen, and T. J. Naughton, "Performance of autofocus capability of deep convolutional neural networks in digital holographic microscopy," (2016).
12. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Commun. ACM **60**, 84 – 90 (2012).
13. Z. Ren, Z. Xu, and E. Y. Lam, "Autofocusing in digital holography using deep learning," in *BiOS,* (2018).
14. Z. Ren, Z. Xu, and E. Y. Lam, "Learning-based nonparametric autofocusing for digital holography," Optica **5**, 337–344 (2018).
15. T. Shimobaba, T. Kakue, and T. Ito, "Convolutional neural network-based regression for depth prediction in digital holography," CoRR **abs/1802.00664** (2018).
16. G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," (2007).
17. M. Tamamitsu, Y. Zhang, H. Wang, Y. Wu, and A. Ozcan, "A robust holographic autofocusing criterion based on edge sparsity: comparison of Gini index and Tamura coefficient for holographic autofocusing based on the edge sparsity of the complex optical wavefront," in *Quantitative Phase Imaging IV,* vol. 10503 G. Popescu and Y. Park, eds., International Society for Optics and Photonics (SPIE, 2018), pp. 22 – 31.
18. L. Ma, H. Wang, Y. Li, and H. Jin, "Numerical reconstruction of digital holograms for three-dimensional shape measurement," J. Opt. A: Pure Appl. Opt. **6**, 396–400 (2004).
19. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," CoRR **abs/1505.04597** (2015).
20. A. Gilles, P. Gioia, R. Cozot, and L. Morin, "Hybrid approach for fast occlusion processing in computer-generated hologram calculation," Appl. Opt. **55**, 5459–5470 (2016).
21. M. Subbarao, T. Choi, and A. Nikzad, "Focusing techniques," J. Opt. Eng. **32**, 2824–2836 (1993).
22. S. Nayar and Y. Nakagawa, "Shape from focus," IEEE Trans. on Pattern Anal. Mach. Intell. **16**, 824–831 (1994).
23. J. Pech-Pacheco, G. Cristobal, J. Chamorro-Martinez, and J. Fernandez-Valdivia, "Diatom autofocusing in brightfield microscopy: a comparative study," in *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000,*

vol. 3 (2000), pp. 314–317 vol.3.
24. A. Thelen, S. Frey, S. Hirsch, and P. Hering, "Improvements in shape-from-focus for holographic reconstructions with regard to focus operators, neighborhood-size, and height value interpolation," IEEE Trans. on Image Process. **18**, 151–157 (2009).
25. G. Yang and B. Nelson, "Wavelet-based autofocusing and unsupervised segmentation of microscopic images," in *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453),* vol. 3 (2003), pp. 2143–2148 vol.3.
26. E. Krotkov and J.-P. Martin, "Range from focus," Proceedings. 1986 IEEE Int. Conf. on Robotics Autom. **3**, 1093–1098 (1986).
27. A. Santos, C. Ortiz-de Solorzano, J. J. Vaquero, J. Peña, N. Malpica, and F. Del Pozo Guerrero, "Evaluation of autofocus functions in molecular cytogenetic analysis," J. microscopy **188**, 264–72 (1998).
28. H. Nanda and R. Cutler, "Practical calibrations for a real-time digital omnidirectional camera," Proc. CVPR, Tech. Sketch (2001).
29. A. Gilles, P. Gioia, R. Cozot, and L. Morin, "Computer generated hologram from multiview-plus-depth data considering specular reflections," 2016 IEEE Int. Conf. on Multimed. & Expo Work. (ICMEW) pp. 1–6 (2016).