PS-NET: an end-to-end phase space depth estimation approach for computer-generated holograms

Nabil Madali, Antonin Gilles, Patrick Gioia, Luce Morin

Research Article Vol. 32, No. 2 / 15 Jan 2024 / Optics Express 2473
Optics EXPRESS

PS-NET: an end-to-end phase space depth estimation approach for computer-generated holograms

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

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

Abstract: In the present work, an end-to-end approach is proposed for recovering an RGB-D scene representation directly from a hologram using its phase space representation. The proposed method involves four steps. First, a set of silhouette images is extracted from the hologram phase space representation. Second, a minimal 3D volume that describes these silhouettes is extracted. Third, the extracted 3D volume is decomposed into horizontal slices, and each slice is processed using a neural network to generate a coarse estimation of the scene geometry. Finally, a third neural network is employed to refine the estimation for higher precision applications. Experimental results demonstrate that the proposed approach yields faster and more accurate results compared to numerical reconstruction-based methods. Moreover, the obtained RGB-D representation can be directly utilized for alternative applications such as motion estimation.

1. Introduction

Computer-generated holography [1] has modernized holography by enabling the creation of highly complex and realistic 3D holograms using numerical techniques. A computer-generated hologram (CGH) can be computed from a synthetic scene by decomposing it into geometric primitives, and then numerically simulating the light diffraction patterns produced at the hologram plane. Unlike traditional images, where each pixel receives information from a single point in the captured scene, in a CGH the light wave scattered by each point in the scene contributes to every pixel during hologram recording. This means that the influence of a single scene point extends to every pixel in the CGH, creating a non-locality that poses challenges for efficient compression [2,3].

The non-locality of CGHs has a substantial impact on the development of precise motion estimation algorithms for holographic video compression [4,5]. This is different from traditional video compression, where motion estimation is usually conducted locally [6] by searching for the most suitable block of pixels within a limited spatial range. In holographic videos, a small motion of the scene can cause changes across large regions of the hologram. In addition, motion estimation in holographic videos requires 3D motion vectors to accurately predict the motion between consecutive frames, as scene motion in the depth axis has an impact on the hologram values, and should thus be estimated for proper motion compensation.

Recently, [7] proposed recovering an RGB-D representation of the scene using the Depth-from-focus (DFF) methods [8–11] given a numerical reconstruction volume computed from the hologram. Then, utilize traditional motion estimation algorithms on the RGB part and estimate additional motion vectors using the computed depth map. The authors extend the approach into learning-based methods by segmenting the in-focus regions using either vertical [12] or horizontal [13] decomposition of the reconstruction volume.

#501085 https://doi.org/10.1364/OE.501085
Journal © 2024 Received 20 Jul 2023; revised 4 Dec 2023; accepted 6 Dec 2023; published 12 Jan 2024

Optics EXPRESS

The proposed numerical reconstruction-based methods have a low inference speed as they heavily rely on numerical reconstruction, and must process a large reconstruction volume that increases linearly with the size of the hologram. Furthermore, the obtained performances are affected by the used pixel pitch; for a larger pixel pitch, the numerical reconstruction is free of speckle noise, and the focus changes at a slower rate, leading to higher accuracy. However, as the pixel pitch decreases, the numerical reconstruction becomes increasingly contaminated by speckle noise, and the focus changes faster, resulting in lower performance.

To alleviate those limitations, an End-to-End approach for recovering RGB-D scene representation using the hologram phase space representation is presented in this paper. The proposed method consists of four steps. First, the hologram phase space representation is computed using the short-time Fourier transform, and the corresponding phase space support is segmented using a neural network. Second, regions of interest (ROI) are extracted using pointwise phase space support intersection. Third, the ROI shapes are then reversed using a neural network to produce a coarse estimation of the scene geometry. Fourth, a third neural network is utilized to refine the estimation for higher precision applications. The proposed approach is highly parallelizable and does not require numerical reconstruction.

The present work aims to design a method that has similar accuracy to the reconstruction-based methods but with faster inference time. Given a sequence of holographic frames, the proposed method can be used to infer an RGB-D representation for each holographic frame. Then, the motion vectors between two consecutive frames can be computed using a classical block-matching approach [6] on the RGB parts, and an additional motion in the z-axis using the predicted depth maps. Finally, the motion-compensated frame can be computed using the segmentation and motion compensation method proposed in [4], by considering each block as a unique object with its own rigid body motion given by the computed 3D-motion vectors.

The remaining of the article is organized as follows: Section 2 reviews previous works that use phase space to process holographic data. Then, Section 3 presents the proposed method. In Section 4, a series of experiments are conducted to validate the proposed approach. Finally, in Section 5, we discuss the advantages and limitations of the proposed approach.

2. Related works

2.1. Numerical reconstruction-based methods

At the present time, the most commonly used technique for extracting scene geometry from input holograms is the depth from focus (DFF) approach [14]. As its name implies, it attempts to determine depth information based on the level of focus, where higher focus levels correspond to closer depths. To apply this method, a 3D volumetric rendering is conducted by numerically propagating the hologram at various manually defined reconstruction distances. Subsequently, the reconstruction volume is spatially divided into either overlapping or non-overlapping patches [15], with each patch considered an independent object. Depth information is then determined by detecting the optimal reconstruction distance at which the focus is highest using a focus measure (FM).

Recently, [7] introduced a novel FM named Patch-based for CGH, aiming to identify and segment the in-focus regions from each numerical reconstruction. The authors initially computed a 3D reconstruction volume within a manually defined reconstruction interval. This obtained volume was then divided into non-overlapping patches of size 32×32, and the results were fed into a U-Net network [16]. The network was supervised using cross-entropy loss to predict the correct in-focus plane at the pixel level. Subsequently, the same authors proposed an alternative approach called H-seg [13], which processes the reconstruction volume horizontally rather than vertically. In this method, the reconstruction volume used in the previous approach was decomposed into horizontal slices, which were then fed into the U-Net network. The network is supervised to segment the in-focus lines present in each horizontal slice. This alternative method

Optics EXPRESS

is designed to alleviate the patch discontinuity present in the patch-based approach. However, it is not well-suited for small recording pixel pitches.

Although these methods give convincing results, they suffer from a high computational cost due to the intensive use of numerical reconstruction. Consequently, in the present study, we propose the use of the hologram phase-space representation (PSR) to devise a more efficient methodology.

2.2. Phase space methods

The PSR of a hologram [17] is a local frequency spectrum representation that enables the semantic interpretation of the hologram contents. Several methods have been proposed in the literature to compute the hologram PSR, including the Wigner-Ville [18] (WV) distribution, Short-term Fourier transform (STFT) [18], and S-method [19]. A comprehensive analysis and visual comparison of these PSR methods are presented and discussed in detail in [20].

The hologram PSR has been used for several applications including fast hologram generation [21], compression [22], segmentation [4], and quality evaluation [23]. However, little attention has been devoted to designing accurate depth estimation algorithms for scenes with complex geometry. Previous works [24–27] on the subject are restricted to microscopic scenes composed of sparse evenly spaced 3D points and rely on the intuition that each 3D point has a unique point spread function (PSF) phase space footprint, which can be easily reversed to estimate the point focal distance. Using this assumption the depth estimation problem can be reformulated as pattern matching, where each hologram phase space representation is decomposed into known PSF phase space footprints, then individually reversed to produce a scene depth estimate.

To effectively decompose the hologram PSR into a set of known PSF patterns, careful consideration must be given to the number of scene points and the displacement between each point. This ensures there is sufficient space between each pattern in the obtained PSR, allowing for easy segmentation.

Recently, [28] have investigated the reversal of PSR of horizontal lines of a computer-generated hologram with complex scene geometry. The authors initially employ a convolutional neural network to extract the support (i.e. non-zero elements) information from the hologram PSR. Subsequently, the obtained support is remapped onto the 3D space to obtain a set of ROI. Each extracted ROI encapsulates an individual scene object with a shape that correlates with the enclosed object points and a PSR support that aligns perfectly with the object PSR support. The authors demonstrated that each ROI shape can be simplified to a single straight line, which maintains the same PSR support as the original shape. This simplified line can serve as a coarse estimate of the scene depth and can be further refined using DFF methods to recover the curvature information.

The present study represents a notable advancement over the previous work by addressing two key limitations. First, the extraction of ROI no longer relies on a set of 1D slices but instead utilizes the 2D hologram directly, which substantially reduces the occurrence of false positive ROI. Second, a neural network rather than an iterative algorithm is employed to reverse the obtained ROI shapes and generate a coarse depth estimate, which allows faster and more accurate results. The network is trained to learn the shape-to-shape relationship between the extracted ROI shape and the encapsulated scene points.

2.3. Shape from silhouette

The proposed approach is founded on a similar idea to the shape-from-silhouette (SfS) method [29], which involves acquiring a set of silhouette images $\{S^i\}$ of a 3D object O using K perfectly calibrated cameras $\{C^i\}$ positioned around it, as illustrated in Fig. 1. Subsequently, the goal is to find a volume $\mathcal{A}$ that encompasses the 3D scene objects and that precisely matches the silhouette

Optics EXPRESS

images, meaning that when projected onto the camera C^i image plane using the associated projection π^i, $\pi^i(\mathcal{A})$ aligns perfectly with the silhouette image S^i for all $i \in \{1, \ldots, K\}$.

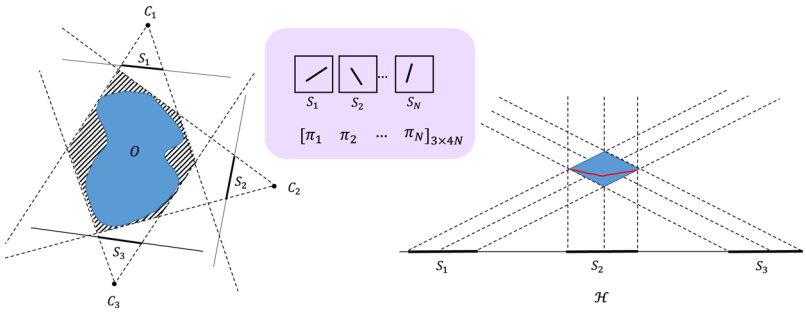
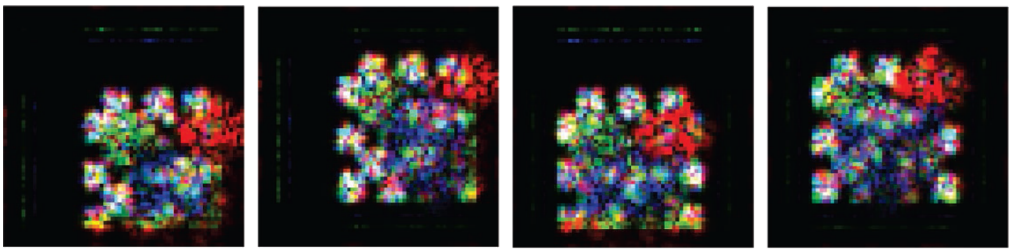


Fig. 1. In the left image, traditional camera settings for the shape-from-silhouette approach, with cameras positioned around the targeted object. In the right image, the camera settings employed in the present study, where the camera centers are positioned along a shared line and represent the position at which the STFT transform is computed.

The accuracy of the captured scene geometry depends on the number of cameras used, their orientations, and the complexity of the scene geometry. A larger number of cameras with carefully chosen orientations leads to a more detailed reconstruction of the geometry. However, employing a large number of cameras can be costly. In such cases, it is often preferable to use a smaller number of cameras and refine the estimated geometry manually using prior knowledge about the processed 3D shape.



(a) (b) (c) (d)

Fig. 2. The obtained local frequency spectrum (i.e. lenslet images) $C[.,.,m_1,m_2]$at different spatial locations (n_1, n_2).

In this study, the silhouette images are not extracted using a pinhole camera. Instead, they are obtained through the binarization of the support structures within the lenslet images provided by the hologram PSR.

The lenslet images extracted from the PSR illustrated in Fig. 2 possess two distinctive characteristics. First, they are limited in resolution, with each image having a size of $M \times M$, where M represents the frequency discretization step. Second, they are positioned on the same plane and not on the viewpoint surrounding the 3D scene (i.e., the hologram plane), making it more challenging to estimate the internal object points accurately. However, the number of accessible silhouette images is on the order of N^2, where N represents the spatial discretization. Hence, we can assume that a specific 3D shape will produce a unique set of silhouette images, and conversely, that these silhouette images will generate a unique volume $\mathcal{A}$ related to the processed shape.

The objective of this work is to design a learning-based approach to extract RGB-D scene representation from the extracted lenslet images.

Optics EXPRESS

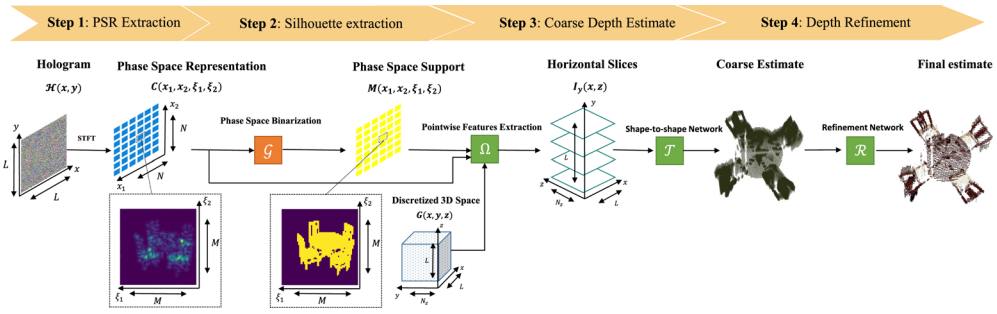


Fig. 3. Illustration of the different steps of the proposed method.

3. Methodology

The proposed method for estimating the scene geometry from hologram PSR is introduced in the following section.

3.1. Overview

The proposed method, as shown in Fig. 3, comprises four main steps. In the first step, the input hologram $\mathcal{H}$ undergoes an STFT transform to obtain its PSR C. This transform can be interpreted as a lenslet representation of the hologram, where each lenslet provides a view of the input 3D scene.

Next, a neural network denoted as $\mathcal{G}$ is utilized to extract the silhouette (i.e. support) S of the lenslet images. Then, the 3D volume with a projection that coincides exactly with the silhouette images is extracted using a pointwise PSR mapping function and then sliced into horizontal slices. Third, a neural network denoted by Γ is used to invert the shape-to-shape relationship between the ROI that intersect the horizontal slices and the 3D scene points, to produce a coarse depth estimate. In the fourth step, a third neural network $\mathcal{R}$ is employed to refine the estimation and improve its accuracy and quality based on color as prior constraints.

3.2. Hologram phase space representation

Given an input hologram denoted by $\mathcal{H} \in \mathbb{C}^{L\times L}$, the first step is to compute the hologram phase space representation denoted C. This is achieved through the application of a windowed STFT using Hann windows of size $M \times M$ with a redundancy factor of $\frac{M}{2}$. The STFT is computed on a grid of uniformly sampled points given, for $i = 1, 2$ by:

$$x_i\,[n_i] = \frac{L\Delta x_i}{2N}(2n_i - N + 1), n_i \in \{0, 1, \ldots, N-1\} \tag{1}$$

and a set of sampled frequencies given by

$$\xi_i[m_i] = \frac{1}{2\Delta x_i M}(2m_i - M + 1), m_i \in \{0, 1, \ldots, M-1\} \tag{2}$$

where Δx_1, Δx_2 are the pixel pitches along the horizontal and vertical directions, and N and M are the space and frequency resolutions, respectively. The complete STFT operation thus provides the 4D PSR C given by,

$$C[n_1, n_2, m_1, m_2] = \sum_{x=0}^{L}\sum_{y=0}^{L} \mathcal{H}[x, y]\, w[x - n_1, y - n_2] e^{-2\pi j(\xi_1[m_1]x + \xi_2[m_2]y)} \tag{3}$$

The PSR can be seen as a set of $N \times N$ lenslet images $C(n_1, n_2))$ of size $M \times M$, each lenslet image being associated with a specific position $(x_1[n_1], x_2[n_2])$.

Optics EXPRESS

3.3. Silhouette extraction

The second step of the method consists of estimating the phase space silhouettes, defined as the support (set of non-null locations) of each lenslet image. This is essentially a binarization of the PSR obtained in the first step. However, a classical binarization process provides bad results due to varying optimal thresholds across different holograms. Moreover, the resulting binary mask tends to be coarse with degraded object outlines and therefore requires additional post-processing steps. A neural network is thus used to ensure fast and accurate PSR binarization.

3.3.1. Ground truth silhouettes

Assuming that the ground truth 3D scene used to generate the hologram $\mathcal{H}$ is available as a 3D point cloud $\mathcal{P}$, the ground truth phase space silhouettes can be estimated by mapping this 3D scene point cloud onto phase space. For a given 3D point p its mapping in the lenslet image $C[n_1, n_2]$ is given by the coordinates (u_1, u_2) defined as,

$$u_i(n_1, n_2) = \text{argmin}_{m_i} \left\{ \left| \hat{\xi}_i(n_1, n_2), \xi_i[m_i] \right| \right\} \tag{4}$$

$$\hat{\xi}_i(n_1, n_2) := -\frac{(x_i[n_i] - y_i)}{\lambda \sqrt{(x_1[n_1] - y_1)^2 + (x_2[n_2] - y_2)^2 + z^2}}, \tag{5}$$

where λ is the used wavelength. The mapping is valid only if the following constraints are fulfilled,

$$-\frac{1}{2\Delta x_i} \leq \xi_i \leq \frac{1}{2\Delta x_i}. \tag{6}$$

Therefore,the PSR support S_p associated with point p can be defined as,

$$S_p[n_1, n_2, u_1(n_1, n_2), u_2(n_1, n_2)] = 1. \tag{7}$$

By summing up individual support for each 3D point p of the 3D scene $\mathcal{P}$, the hologram PSR can be obtained:

$$S = (\sum_{p \in \mathcal{P}} S_p) > 0. \tag{8}$$

The obtained binary mask S provides the location of non-null samples for this specific 3D scene in the hologram PSR, that is the ground truth silhouettes.

3.3.2. Neural network-based extraction

A neural network denoted by $\mathcal{G}$ based on the U-Net architecture [16] is used to estimate a phase space support $\hat{S}$ that is as close as possible to the ground truth S given as input the PSR C, more formally:

$$\hat{S} = \mathcal{G}_{\theta_1}\{(\mathbb{R}e\{C\}, \mathbb{I}m\{C\})\}, \quad C \in \mathbb{C}^{N^2 \times M \times M} \tag{9}$$

$$\min_{\theta_1} \mathcal{L} = BCE(\hat{S}, S), \quad S, \hat{S} \in \mathbb{R}^{N^2 \times M \times M} \tag{10}$$

where $\mathbb{R}e\{C\}, \mathbb{I}m\{C\}$ are the real and imaginary parts of C, θ are the weights of the network, and BCE is the binary cross-entropy loss.

3.4. Coarse depth estimate

In the third step, a coarse depth estimate is computed as a binary occupancy volume, by selecting the 3D points that map onto phase space inside the estimated silhouette $\hat{S}$, as in the classical SfS algorithm.

Optics EXPRESS

3.4.1. Phase space support intersection score

The 3D space is discretized into a uniformly sampled grid of points defined by $\mathcal{V} = X^1 \times X^2 \times X^3$, with

$$X^j = \left\{ \frac{i \cdot (x^j_{min} - x^j_{max})}{N^i - 1} + x^j_{min} : \quad i = 0, \ldots, N^j - 1 \right\} \tag{11}$$

where N^i is the number of sampled points and x^j_{min}, x^j_{max} define the sampling boundary. Then the 3D volume $\mathcal{V}$ is parsed and each 3D point support S_p is intersected with the estimated silhouette $\hat{S}$ as follows:

$$s_p = \frac{S_p \circ \hat{S}}{\sum S_p}, \tag{12}$$

where $\circ$ is an element-wised dot product between the two matrices.

In the case of a perfect silhouette estimation, all the points belonging to the scene will have a score equal to one. However, in practice, this ratio will be lower, as the silhouette estimation is not perfect. In addition, due to the limited phase space resolution, multiple points surrounding the scene points will have a score of one, forming an ROI that encompasses the scene points. These ROIs are shaped in correlation with the enclosed scene points, as stated in [28]. In simple cases, when the scene points form straight lines with a specific slope, the ROI shape can be optimally recovered. However, when more intricate shapes are present, merely using shape information to reverse the ROI shape will produce suboptimal results. This is why a piece of additional color information and a second refinement step with spatial depth constraints are added to generate a more accurate depth estimate.

3.4.2. Phase space color

To improve the estimated depth, the value of the phase space coefficients impacted by a 3D point p are considered, and not only their non-nullity. The phase space coefficients associated with the point p are given as:

$$\Phi(p) = \left\{ c \in C \circ S_p \ : \ \text{if } c \neq 0 \right\} \tag{13}$$

The cardinal of the set $\Phi(p)$ varies according to the position of the point p with a maximum cardinal equal to $N \times N$, and each of its coefficients should depend on the color of p. However, because of limited phase space resolution, multiple 3D points may have common coefficients among their set of associated coefficients, leading to phase space coefficients with a mixture of colors from different 3D objects. In this study, we assume that values of coefficients in $\Phi(p)$ are distributed around a mean value corresponding to the color of the 3D point p. Formally, the color c_p of the 3D point p is estimated as follows:

$$c_p = f_c \left(\frac{1}{\mid \Phi(p) \mid} \sum_{c \in \Phi(p)} c \right), \tag{14}$$

where f_c is defined as a simple multilayer perceptron (MLP) composed of one hidden layer with 128 hidden neurons.

Figure 4 demonstrates that the mean of the phase space coefficients captures the 3D point color information, albeit some loss in contrast that can be recovered using the mapping function f_c. However, the obtained amplitude map is blurry and lacks details compared to the result computed from the numerical reconstruction. This is due to the limited phase space resolution.

In order to better understand the relation between the phase space resolution and the color information, an epipolar image is illustrated in Fig. 5. The epipolar image is computed by fixing one pair of spatial and frequency terms (n_2, m_2) in the original PSR representation, namely, $C[., n_2, ., m_2]$. As depicted in Fig. 5, the hologram PSR limited resolution causes the color

Optics EXPRESS

(a) (b) (c) (d)

Fig. 4. The figure illustrates the ground truth amplitude map in (a), the amplitude map computed from numerical reconstruction in (b), the amplitude map computed from the mean of the phase space coefficients in (c), the amplitude map finally computed after applying function f_c in (d).

information of multiple points to blend, making it challenging to extract individual points solely based on the color information of the epipolar lines. Consequently, the computed color information may not display significant color change between adjacent points along the XZ plane. Nevertheless, this information can still assist the network in distinguishing between multiple objects within the same ROI and independently processing the shape of each object.

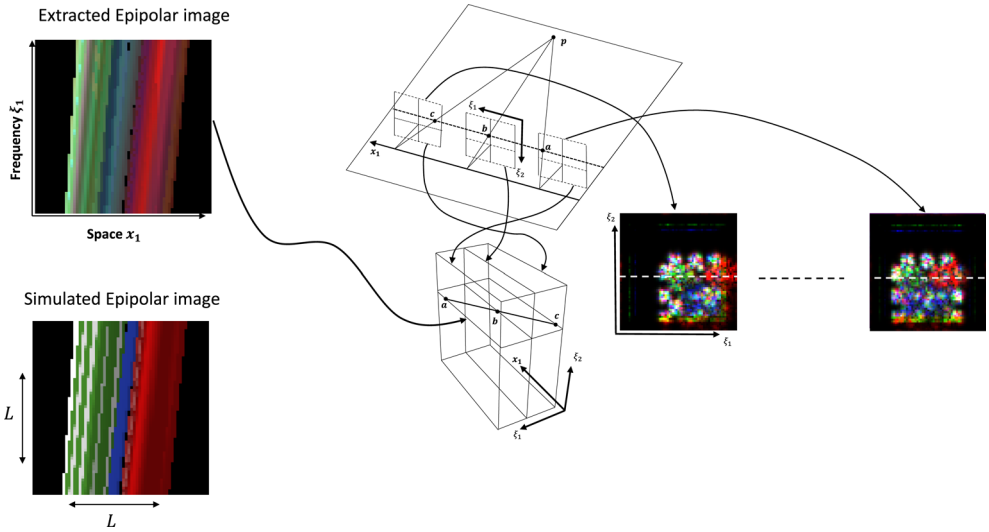


Fig. 5. The figure illustrates the geometrical interpretation of the extracted epipolar image $C[n_1, ., m_1, m_1]$ from the hologram PSR. In the top left of the image is the extracted epipolar image, and in the bottom left is the expected epipolar image if the PSR resolution was not limited. The simulated image is generated by projecting the scene point cloud using the projection formula in Eq. (5), where both N and M are set equal to L.

3.4.3. Depth map extraction

Finally, a coarse depth map is estimated based on the extracted discretized 3D volume with the associated intersection score s_p and color information c_p of each 3D discretized point p. To do so, the volume $\mathcal{G}$ is decomposed into horizontal slices I_y with elevation y defined as,

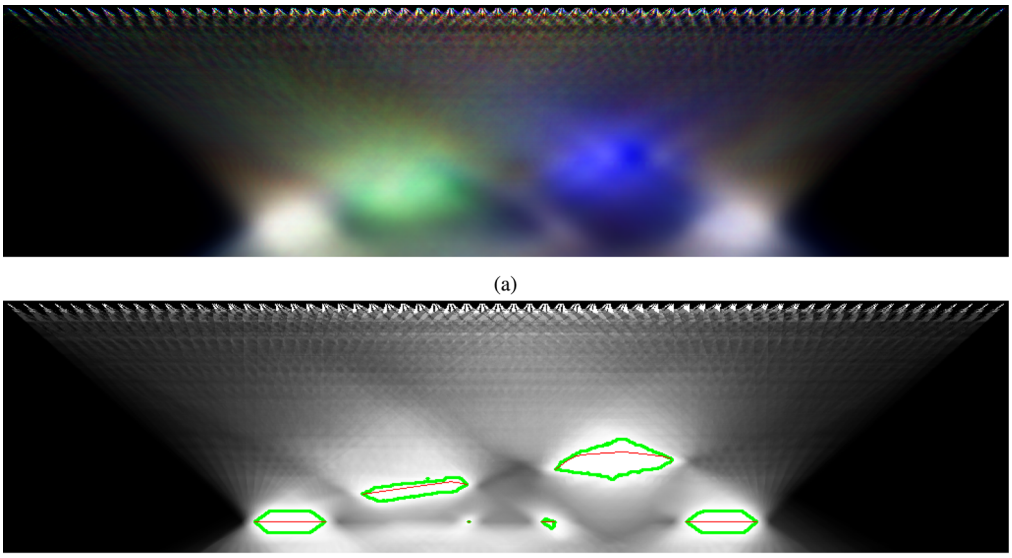
$$I_y(x, z) = [s_p; c_p] \quad p = (x, y, z) \in \mathbb{R}^3. \tag{15}$$

Optics EXPRESS

Figure 6 illustrates an example of an extracted horizontal slice, along with its corresponding intersection scores and color information. Points with an intersection score equal to one define the ROI patterns, whose border is highlighted in green on the figure. These patterns emerge from the diffraction of light rays emitted from the hologram towards the scene points. Each ROI encapsulates scene points illustrated in red, with a shape that correlates with the shape encapsulated scene points.

In this study, instead of relying on manual filtering of the obtained score s_p to retain only points with a score of 1, which may lead to a degradation in the regions of interest shape. A neural network is used to extract both the ROI boundaries and the coarse depth estimate from the intersection scores.

In summary, each horizontal slice is processed by a neural network denoted $\mathcal{T}$ defined as Deep Residual U-Net [30], that utilizes shape and color information provided by the horizontal slice coefficients values to estimate coarse scene depth $\hat{d}_y$ which is the closest possible to the ground truth depth d_y, as illustrated in Fig. 7. The network architecture illustrated in Fig. 8 is composed of two distinct paths: a contracting path and an expansion path. The contracting path is defined as the first four residual blocks of the ResNet50 [31] architecture and is utilized to extract a multi-scale representation of the input feature space at scales 1/2,1/4,1/8,1/16. The expansive path is the same as that used in the original U-Net paper [16], with only the replacement of the bilinear sampling operation by a 2D-transposed convolution.



(a)

(b)

Fig. 6. The figure illustrates both the color information in (a) and the shape information represented by the intersection score and the ROI in green in (b).

This process is more formally described as

$$\hat{d}_y = \mathcal{T}_{\theta_2}\{I_y\} \tag{16}$$

$$\min_{\theta_2} \mathcal{L} = BCE(\hat{d}_y, d_y) \tag{17}$$

where θ_2 is the network parameters, and d_y is the ground truth depth map at elevation y. The final coarse depth $\bar{D}$ and amplitude $\bar{A}$ maps are given as:

$$\bar{D}(x, y) = \text{argmax}_z \hat{d}_y(x, z), \tag{18}$$

Optics EXPRESS

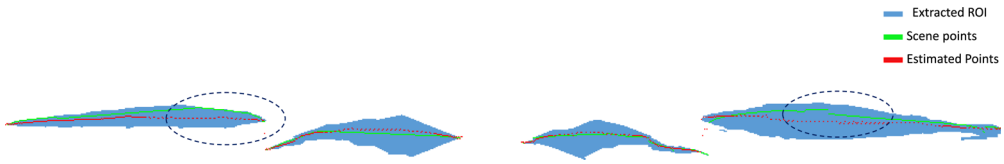


Fig. 7. An example of an inverted shape-to-shape relationship using the pre-trained neural network. The predicted points are relatively close to the ground truth points, however, a refinement step is necessary for highly curved objects to improve the obtained predictions at the inner object points.

$$\bar{A}(x, y) = c(p), \quad p = (x, y, \bar{D}(x, y)). \tag{19}$$

3.5. Depth refinement

The coarse depth estimate is generated without considering the spatial correlation between different horizontal slices and tends to have a better estimate at the borders of objects where the visual cones highly intersect, contrarily to objects inner points. In this fourth and last step, a third neural network $\mathcal{R}$ based on the U-Net architecture is used to refine the coarse depth estimate $\bar{D}$ by ensuring the spatial coherence of the produced depth map guided by the predicted amplitude map $\bar{A}$, more formally:

$$\hat{D} = \mathcal{R}_{\theta_3}\{[\bar{D}; \bar{A}]\} \tag{20}$$

$$\min_{\theta_3} \mathcal{L} = MSE(\hat{D}, D) \tag{21}$$

where D is the ground truth depth map, θ_3 is the network parameters, and MSE is mean squared error loss.

4. Experiments

4.1. Experimental setup

The proposed approach is trained and evaluated using the *Open-access IRT b-com datasets* [32]. The dataset consists of six distinct scenes: *Piano*, *Table*, *Woods*, *Cars*, *Dices*, and *Animals*. Each scene includes 300 frames captured along predefined paths. For every frame, the dataset provides the corresponding recorded hologram, along with the associated depth and amplitude information. The phase information for each holographic frame is randomly generated during the hologram recording, and it is not included in the actual datasets. These depth and amplitude values serve as ground truth in the experimental setup. The first four scenes were used for training and validation, while the remaining two were exclusively for testing. The test and validation sets were sampled every ten frames starting from the first frame, and the remaining 270 frames were used for the training set. Since each hologram is of size 1024×1024 with a pixel pitch of 1μm, and the z axis is discretized into 256 values, the total number of horizontal slices for shape-to-shape network $\mathcal{T}$ training is $1024 \times 270 \times 4$. During the training process, each possible horizontal slice is randomly flipped along the horizontal or vertical axis to increase the size and diversity of the dataset, which prevents overfitting and increases the generalization ability of the model. The same data augmentation technique is used on the 270×4 predicted coarse depth maps to train the refinement network $\mathcal{R}$.

The phase space binarization $\mathcal{G}$ network pre-trained in [28] is re-used in the present experiments to infer the phase space support given the hologram's phase space representation.

Both the coarse and refinement networks are trained for 200 epochs using the stochastic gradient descent and cyclic learning rate decay between 0.1 to 0.01, and early stopping if the

Optics EXPRESS

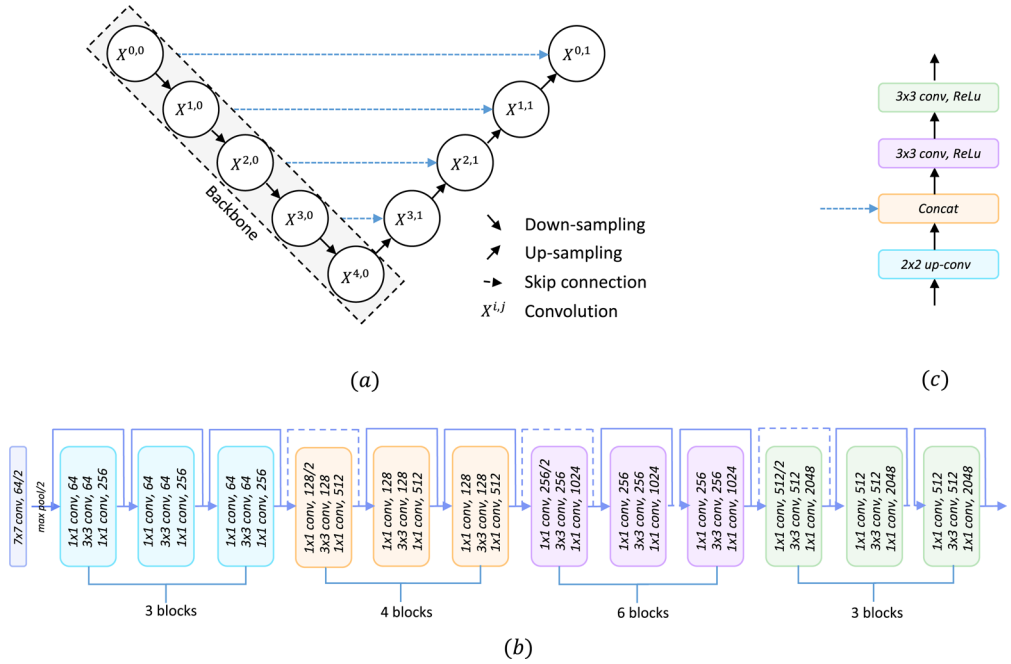


Fig. 8. Illustration of the network architecture in (a), in (b) is the backbone ResNet50 network, and in (c) is the Up-sampling block.

validation loss didn't decrease for 10 epochs. The networks undergo two rounds of training and evaluation. In the first round, the ground truth phase space supports are used, while in the second round, the predicted supports are used. This approach allows us to assess the impact of silhouette extraction on the final results.

The proposed approach is compared to two state-of-the-art approaches namely the patch-based [7] and the H-seg [13] methods, using the ℓ_1 norm between the ground truth and the predicted depth map as performance metrics. The obtained ℓ_1 norm corresponds to the distance between the prediction and the actual ground truth focus plane, expressed as the focal plane indexes difference.

4.2. Results

Table 1 gives the obtained validation and test set results, and a visual results are provided in Figs. 9–11.

First, the table shows that for methods based on retrieving depth information through in-focus detection using hologram numerical reconstruction volumes, the patch-based method performs better than the H-seg approach. This difference in performance can be attributed to the small pixel pitch used for hologram recording, which results in a rapid focus change between consecutive numerical reconstructions. As a result, the in-focus lines that the H-seg method targets are too fine to be clearly visible on the inputted horizontal slices, making it difficult for the network to rely only on local features designed for focus change detection, as would be the case with larger pixel pitches. Instead, the network must rely on global features to describe the complex process of in-focus line formation.

The patch-based method, on the other hand, employs a vertical reconstruction volume crop. This technique ensures that even if there is a rapid change in focus, there will always be one or more images capturing in-focus details at different depths. Consequently, the network can focus

Optics EXPRESS

Table 1. L1 error using different approaches for validation and test sets.

	Piano	Table	Woods	Cars	Dices	Animals
Numerical reconstruction						
Patch-based	8.28	11.94	5.26	4.33	5.5	4.49
H-seg	9.17	12.9	6.95	8.98	9.23	9.59
Phase Space						
			Ground Truth Support			
PS-Net Coarse	3.05	1.67	1.68	2.51	1.42	1.62
PS-Net Refined	**0.45**	**0.43**	**0.56**	**0.58**	**0.42**	**0.34**
			Predicted Support			
PS-Net Coarse	3.78	5.04	5.39	5.31	3.64	3.55
PS-Net Refined	1.15	1.95	1.49	1.85	1.18	1.63

solely on learning local features that describe in-focus regions, a relatively easier function to learn. Both networks give poor results in the *Table* scene which contains several occluded areas and where the objects are relatively close to each other.

Second, the proposed method coarse estimate outperforms both numerical reconstruction methods on both the test and validation sets when using ground truth phase space support masks. In particular, the highest ℓ_1 norm was achieved in the *Piano* scene, indicating that the learned network $\mathcal{T}$ is capable of accurately reversing the shape-to-shape relationship between the obtained ROI shapes and the enclosed scene points. However, the predicted estimate still requires refinement to recover curved shapes that are prevalent in the *Piano* scene.

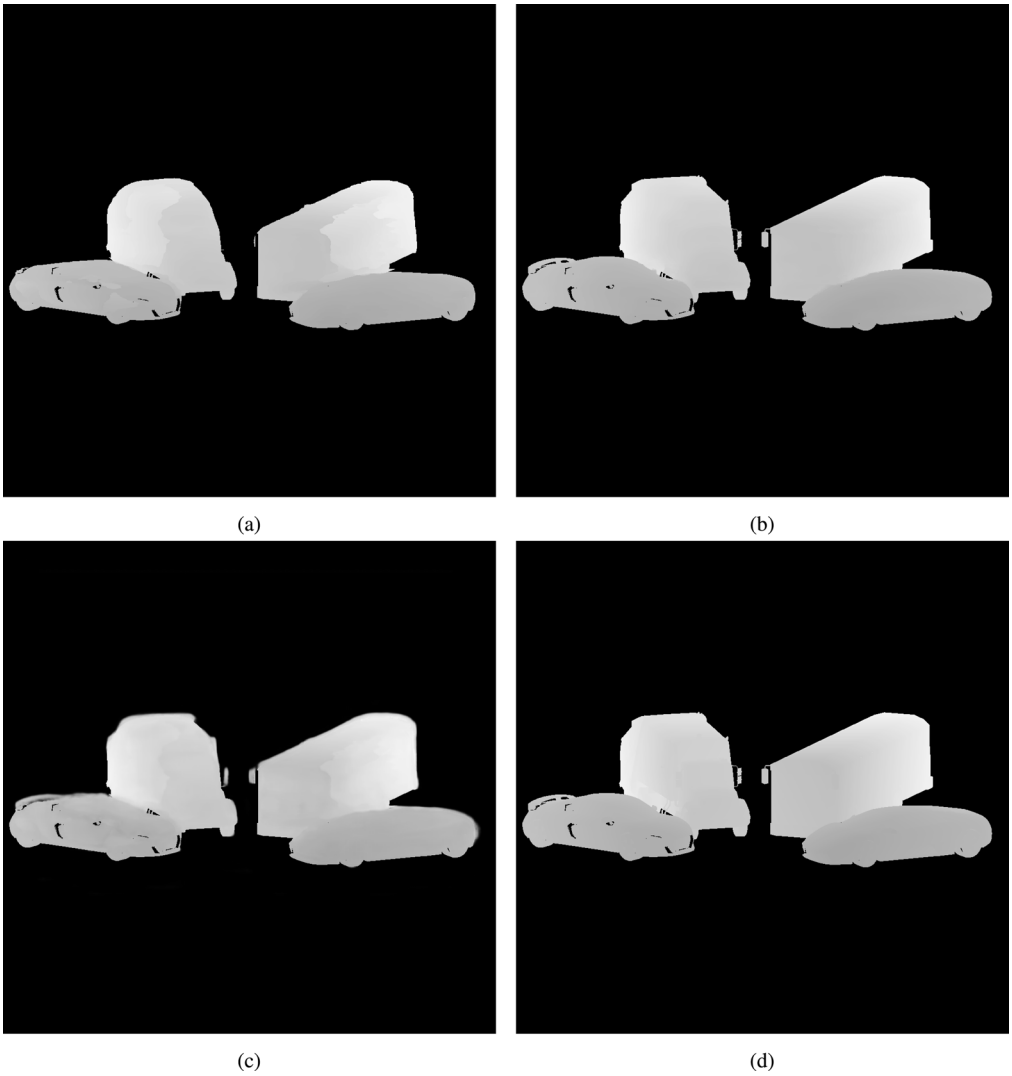
When comparing the obtained results to those predicted using segmented support, we observe that the ℓ_1 norm is multiplied by a factor of two. However, for the *Piano* scene, the results are relatively close, which can be attributed to the fact that there is only one object in the scene. As a result, the phase space support is concentrated in a smaller area, making binarization easier and resulting in less degraded ROI. In contrast, scenes with multiple objects that are spatially separated are more challenging to segment accurately, which can lead to greater errors and lower quality ROI.

Figure 9 provides a visual comparison between the coarse estimate produced using the ground truth and the segmented phase space support. It is evident from the Figure that the coarse estimate using the ground truth support generates an accurate estimate that closely matches the ground truth. In contrast, the estimate produced using the predicted mask is more degraded, with sharp depth transitions between different parts of the objects. Although the refined depth map smoothens the obtained depth map, there are still some areas that are misclassified.

The segmented support can degrade unpredictably based on the scene being processed, the extracted ROI are thus degraded according to the obtained binarization. Therefore the shape-to-shape network $\mathcal{T}$ must extract various ROI shapes from each horizontal slice while accounting for the fact that multiple damaged shapes may correspond to a single enclosed scene points shape. This means that the network has to learn to generalize across multiple inputs that produce the same output, which can be difficult without additional constraints or prior knowledge. The network may struggle to learn an accurate mapping without a proper understanding of the relationship between the inputs and outputs, leading to degraded performance. The additional color information, which provides similar diffraction patterns to ROI and without being subject to a random degradation, can help the network converge to the appropriate function $\mathcal{T}$.

Third, the refined estimates significantly improved the obtained ℓ_1 in both cases. As depicted in Fig. 10, the performance gain can be attributed to two key factors. First, the network incorporates both local and global information to smooth areas with depth discontinuities. Second, it utilizes

Optics EXPRESS



(a) (b)

(c) (d)

Fig. 9. The figure illustrates the obtained depth map using the ground truth in (b) and the predicted phase space in (a). The refined coarse estimate in (c) of the depth map in (a), and the ground truth depth map in (d).

Optics EXPRESS

the amplitude map to recover regions that were initially misclassified, by propagating depth values from the well-estimated areas to poorly estimated areas. However, the amplitude map has inherent limitations. For instance, if a region is poorly estimated in the initial estimate, it will appear blurred in the amplitude map, and thus provide inadequate color information for guiding the depth value propagation by the network. This limitation is especially noticeable when dealing with occluded objects that are adjacent to each other. In such cases, the computed amplitude map will not provide sharp separation boundaries therefore making it difficult to properly separate the two objects in the resulting depth map.

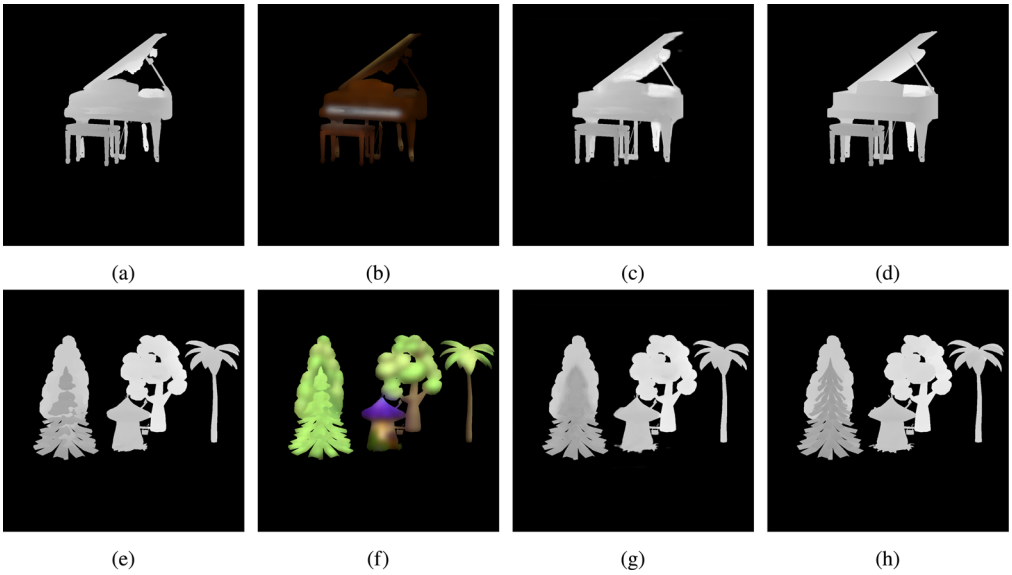


Fig. 10. Each row contains a sample from the test set (Piano and Woods), and illustrates the coarse estimate in (a),(e), the predicted amplitude map in (b),(f), the refined estimate in (c),(g), and the ground truth depth map in (d),(h).

The performance gap between the numerical reconstruction methods and the proposed approach stems from the pre-localization of scene objects in 3D space enabled by the ROI. By narrowing the network's focus to a restricted area of space, the ROI facilitates the generation of a coarse estimation of the scene geometry. This estimation is inherently proximal to the scene points, as it is derived from the ROI, which itself encapsulates the scene points. However, as the pixel pitch increases, spatial localization becomes less efficient, and it may be more appropriate to use numerical reconstruction-based methods for accurate depth estimation.

4.3. Computation time

For each hologram, the computation of the 3D numerical reconstruction volume took approximately 6 minutes to complete with ASM implemented using the native 2021 Matlab Fourier transform implementation on an 11th Gen Intel Core i9-11900F. Then, given the computed reconstruction volume, the patch-based method takes 26.05 seconds to process, while the H-seg method takes 14.30 seconds.

In the proposed method, the phase space segmentation takes approximately 5.3 seconds to complete. The computation of phase space intersection scores and color requires about 2.6 minutes. The inference time for the shape-to-shape network is consistent with that of the H-seg method, standing at 14.30 seconds. Additionally, the refinement network operates with an inference time of 0.8 seconds.

Optics EXPRESS

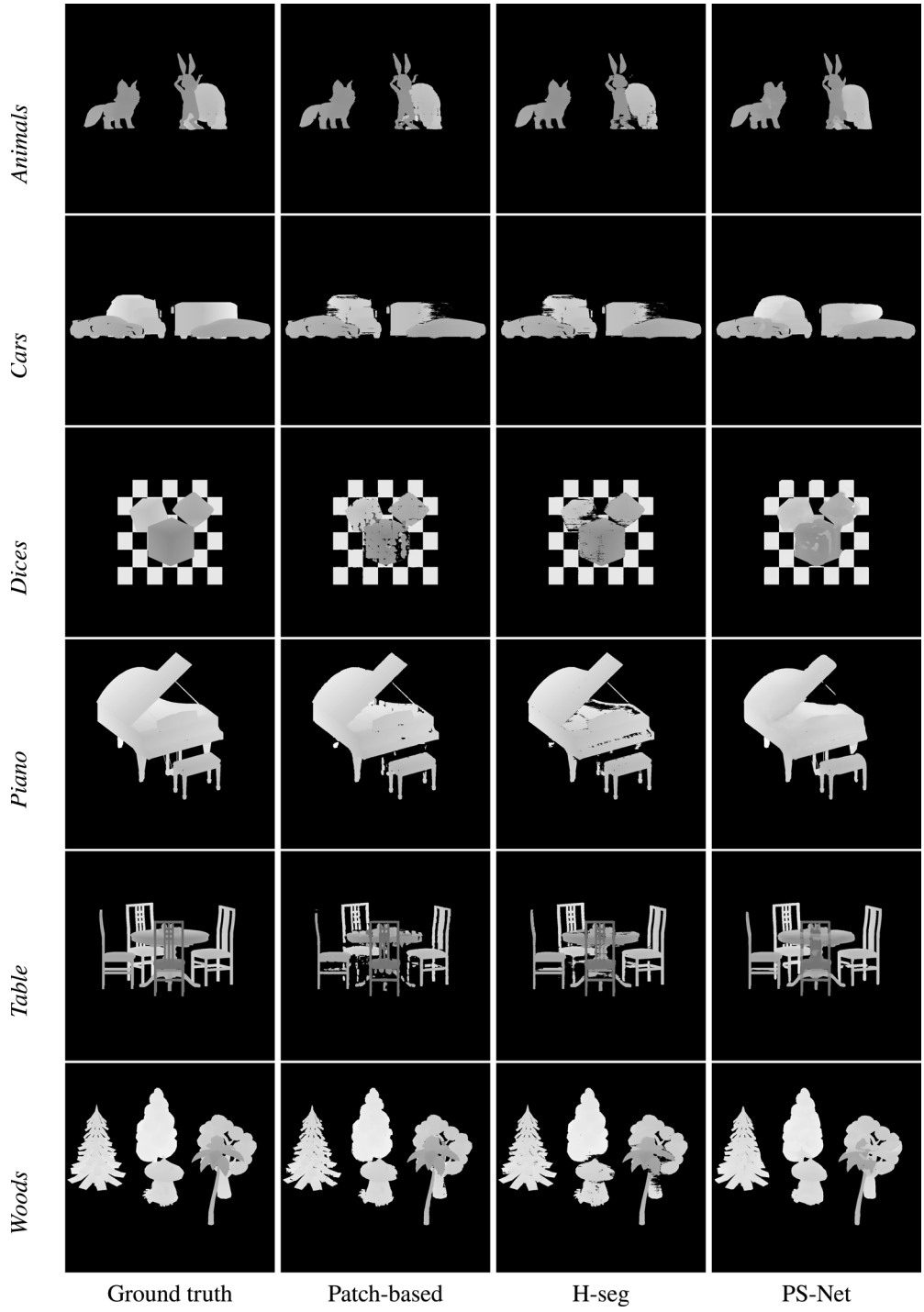


Fig. 11. The obtained depth maps for the different evaluation methods.

Optics EXPRESS

The computation of color and intersection scores can be significantly reduced by first segmenting the central view from the phase space representation $C[N/2, N/2, ., .]$, then upscaling the obtained results to the same resolution as the targeted depth. Finally, only considering the points that are segmented as foreground points as candidates. This step directly defines the spatial boundary of the scene points, and it remains to define their corresponding axial points.

In summary, the proposed approach offers superior speed and accuracy compared to numerical reconstruction methods. However, the accuracy of the results is dependent on the quality of the support binarization, which can be a challenging task in complex scenes with highly curved objects.

5. Conclusion

This paper presents an end-to-end approach for recovering scene geometries from holographic data, using a four-step process. First, the hologram phase space representation is extracted and the corresponding phase space support is segmented. Second, ROI are extracted. Third, a neural network is used to estimate the coarse scene geometry by reversing the shape-to-shape relationship between the ROI shape and the shape of the encapsulated scene points in each ROI. Fourth, a third neural network refines the coarse estimate by ensuring spatial depth consistency along the horizontal and vertical axes.

Experimental results demonstrate that our approach provides faster and more accurate results than numerical reconstruction-based methods. However, the performance of the proposed method is dependent on the quality of the phase space support, which may be degraded for complex scenes.

Future efforts will be directed toward enhancing the results by integrating spatial correlations directly into the shape-to-shape network. Furthermore, the refinement method could undergo a redesign to improve the depth estimate by leveraging both the initial estimate and the 3D discretized space with color information, similar to the patch-based approach. This strategy would facilitate the recovery of occluded regions and lead to additional enhancements.

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. Pi, J. Liu, and Y. Wang, "Review of computer-generated hologram algorithms for color dynamic holographic three-dimensional display," Light: Sci. Appl. **11**(1), 231 (2022).
2. R. K. Muhamad, T. Birnbaum, D. Blinder, *et al.*, "Interfere: A unified compression framework for digital holography," in *Digital Holography and 3-D Imaging 2022*, (Optica Publishing Group, 2022), p. Th4A.2.
3. P. Schelkens, A. Ahar, A. Gilles, *et al.*, "Compression strategies for digital holograms in biomedical and multimedia applications," Light: Adv. Manuf. **3**(3), 1 (2022).
4. T. Birnbaum, D. Blinder, R. K. Muhamad, *et al.*, "Object-based digital hologram segmentation and motion compensation," Opt. Express **28**(8), 11861–11882 (2020).
5. J. P. Q. Peixeiro, "Holographic Information Coding," Master's thesis, Instituto Superior Técnico (2016).
6. W. Hassen and H. Amiri, "Block matching algorithms for motion estimation," in *2013 7th IEEE International Conference on e-Learning in Industrial Electronics (ICELIE)*, (2013), pp. 136–139.
7. N. Madali, A. Gilles, P. Gioia, *et al.*, "Automatic depth map retrieval from digital holograms using a depth-from-focus approach," Appl. Opt. **62**(10), D77–D89 (2023).
8. S. Pertuz, D. Puig, and M. A. Garcia, "Analysis of focus measure operators for shape-from-focus," Pattern Recognit. **46**(5), 1415–1432 (2013).
9. T. Colomb, N. Pavillon, J. Kühn, *et al.*, "Extended depth-of-focus by digital holographic microscopy," Opt. Lett. **35**(11), 1840–1842 (2010).
10. W. Chen and C. Tay, "Extended depth of focus in a particle field measurement using a single-shot digital hologram," Appl. Phys. Lett. **95**(20), 201103 (2009).
11. S. Cuenat, L. Andréoli, A. N. André, *et al.*, "Fast autofocusing using tiny transformer networks for digital holographic microscopy," Opt. Express **30**(14), 24730–24746 (2022).

Optics EXPRESS

12. N. Madali, A. Gilles, P. Gioia, *et al.*, "Automatic depth map retrieval from digital holograms using a deep learning approach," Opt. Express **31**(3), 4199–4215 (2023).
13. N. Madali, A. Gilles, P. Gioia, *et al.*, "H-seg: a horizontal reconstruction volume segmentation method for accurate depth estimation in a computer-generated hologram," Opt. Lett. **48**(12), 3195–3198 (2023).
14. P. Ferraro, S. Grilli, D. Alfieri, *et al.*, "Extended focused image in microscopy by digital holography," Opt. Express **13**(18), 6738–6749 (2005).
15. C. P. McElhinney, B. M. Hennelly, and T. J. Naughton, "Extended focused imaging for digital holograms of macroscopic three-dimensional objects," Appl. Opt. **47**(19), D71–D79 (2008).
16. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," arXiv, arXiv:1505.04597 (2015).
17. J. Ojeda-Castaneda, M. Testorf, and B. Hennelly, *Phase-Space Optics: Fundamentals and Applications: Fundamentals and Applications* (McGraw-Hill Education, 2009).
18. L. Stankovic, *Digital Signal Processing: With Selected Topics: Adaptive Systems, Time-Frequency Analysis, Sparse Signal Processing* (CreateSpace Independent Publishing Platform, 2015).
19. L. Stankovic, "A method for time-frequency analysis," IEEE Trans. Signal Process. **42**(1), 225–229 (1994).
20. T. Birnbaum, T. Kozacki, and P. Schelkens, "Providing a visual understanding of holography through phase space representations," Appl. Sci. **10**(14), 4766 (2020).
21. D. Blinder and P. Schelkens, "Accelerated computer generated holography using sparse bases in the stft domain," Opt. Express **26**(2), 1461–1473 (2018).
22. T. Birnbaum, A. Ahar, D. Blinder, *et al.*, "Wave atoms for digital hologram compression," Appl. Opt. **58**(22), 6193–6203 (2019).
23. G. Finke, M. Kujawińska, and T. Kozacki, "Visual perception in multi slm holographic displays," Appl. Opt. **54**(12), 3560–3568 (2015).
24. L. Onural and M. T. Özgen, "Extraction of three-dimensional object-location information directly from in-line holograms using wigner analysis," J. Opt. Soc. Am. A **9**(2), 252–260 (1992).
25. M. T. Özgen and K. Demirbaş, "Cohen's bilinear class of shift-invariant space/spatial-frequency signal representations for particle-location analysis of in-line fresnel holograms," J. Opt. Soc. Am. A **15**(8), 2117–2137 (1998).
26. S. Oh, C.-Y. Hwang, I. K. Jeong, *et al.*, "Fast focus estimation using frequency analysis in digital holography," Opt. Express **22**(23), 28926–28933 (2014).
27. G. Rajshekhar, S. S. Gorthi, and P. Rastogi, "Estimation of the phase derivative using an adaptive window spectrogram," J. Opt. Soc. Am. A **27**(1), 69–75 (2010).
28. N. Madali, A. Gilles, P. Gioia, *et al.*, "Psdfh: A phase-space-based depth from hologram extraction method," Appl. Sci. **13**(4), 2463 (2023).
29. G. K. M. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette across time part i: Theory and algorithms," Int. J. Comput. Vis. **62**(3), 221–247 (2005).
30. Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual u-net," arXiv, arXiv:1711.10684 (2017).
31. K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," arXiv, arXiv:1512.03385 (2015).
32. A. Gilles, P. Gioia, N. Madali, *et al.*, "Open access dataset of holographic videos for codec analysis and machine learning applications," in *2023 15th International Conference on Quality of Multimedia Experience (QoMEX) (QoMEX 2023)*, (Ghent, Belgium, 2023), p. 6.