H-Seg: a horizontal reconstruction volume segmentation method for accurate depth estimation in a computer-generated hologram

Nabil Madali, Antonin Gilles, Patrick Gioia, Luce Morin

H-Seg: A Horizontal Reconstruction Volume Segmentation Method for Accurate Depth Estimation in Computer-Generated Hologram

Nabil Madali[1,3*] Antonin Gilles[1] Patrick Gioia[1,2] Luce Morin[1,3]

[1] IRT b<>com Cesson-Sévigné France
[2] Orange Labs Cesson-Sévigné France
[3] INSA Rennes Rennes France

Abstract In this work, we introduce a novel approach for depth estimation in CGH by employing horizontal segmentation of the reconstruction volume instead of conventional vertical segmentation. The reconstruction volume is divided into horizontal slices and each slice is processed using a residual U-net architecture to identify in-focus lines, enabling the determination of the slice's intersection with the 3D scene. The individual slice results are then combined to generate a dense depth map of the scene. Our experiments demonstrate the effectiveness of our method, with improved accuracy, faster processing times, lower GPU utilization, and smoother predicted depth maps compared to existing state-of-the-art models.

Accurate and fast motion estimation between two consecutive holographic video frames remains a challenge in the current state of the art due to the hologram signal properties [1]. Unlike natural images where scene objects are well localized inside the image plane, in holography, the light wave scattered by each scene point contributes to every pixel during hologram recording. As a result, the scene objects are poorly localized inside the hologram plane, preventing the use of traditional optical flow methods [2] based on block matching to estimate the scene motion vectors. In addition to the spatial displacement along the horizontal and vertical axis, the motion vector along the optical axis also needs to be estimated. To this end, the scene geometry must first be extracted from the hologram, then motion vector estimation can be performed from this new representation. The Depth-From-Focus (DFF) [3] method is a widely researched technique in digital holographic microscopy for capturing relatively flat scenes, which can be represented by a limited number of focus planes. Unlike CGH, which faces the challenge of varying depth values between adjacent pixels and thus necessitates per-pixel estimation. The recent advances in the subject are discussed in [4, 5, 6].

Recently, [7] explored the use of the DFF method for extracting RGB-D representations from computer-generated holograms. This is achieved by computing numerical reconstructions at uniformly spaced distances within a specified range, retaining only the amplitude information to form the reconstruction volume. The focus level of each pixel is determined using a focus measure operator on a centered patch, and the depth at which the focus is at its maximum is selected as the pixel's depth estimate. The pixel's intensity value is then linked to the holographic reconstruction performed at the estimated depth. The experimental results showed that the DFF method gives reasonable results when the reconstruction distances, patch size, and focus measures are well chosen. However, the method is computationally expensive and thus cannot be used in real-time processing.

The authors extended their work in [8], using a CNN model to evaluate the focus level of each pixel in the reconstruction volume. To ensure that two points cannot share the same horizontal and vertical coordinates and have different depths, the network was fed with a cropped reconstruction volume and was supervised using the Cross-entropy to maximize the focus value at the optimal focus plane. During the inference, the argmax value was taken along the batch axis to predict a local depth map. Unlike DFF where the focus is evaluated at the pixel level using a centered patch, the method proposed in [7] uses a non-overlapping reconstruction volume decomposition and therefore ensures a faster inference time. In addition, the experimental results showed that the proposed approach is more accurate and produces better depth maps than the classic DFF methods. The pre-trained network can be particularly useful for inferring the scene geometry when the ground truth depth map is not available. This situation may arise, for example, when dealing with optically acquired or synthetic holograms that are derived from multi-view or light-field data.

Even though the above-mentionned approach gives promising results, it has some limitations. First, the GPU consumption is linear with the number of holographic reconstructions and the patch size. The higher the number of reconstructions and the bigger the patch size, the greater the 3D volume at the entrance of the network. Second, the network performances are poor

*This work has been achieved within the Research and Technology Institute b<>com, dedicated to digital technologies. It has been funded by the French government through the National Research Agency (ANR) Investment referenced ANR-A0-AIRT-07. Authors can be reached at {nabil.madali, antonin.gilles, patrick.gioia, luce.morin}@b-com.com.

along the border of the patches, thus creating discontinuities in the final depth map. Finally, similarly to DFF methods, the resulting depth map requires the computation of a binary mask to segment the foreground and background objects, resulting in additional computational costs.

In this paper, we propose to reformulate the depth estimation problem, which is typically solved by fixing a 2D region along a plane parallel to the hologram and detecting the reconstruction distance at which the focus is sharper. Here, we propose to perform the depth prediction along horizontal planes perpendicular to the hologram, to distinguish optimal focus planes more easily due to holographic reconstruction characteristics. In contrast to the patch-based approach, the following approach has a lower GPU consumption; additionally, by using the entire horizontal plane at the network input, it is easier to maintain depth continuity, resulting in a smoother depth map. Finally, the predicted depth map does not require additional segmentation into foreground and background classes.

The proposed approach for retrieving the scene geometry from a given hologram is depicted in Figure 1. The process begins by building a reconstruction volume using a series of numerical reconstructions. Next, the volume is broken down into a set of horizontal slices, and a CNN model is trained to estimate the in-focus lines present on each slice. The final phase of the process involves refining the segmentation results to eliminate occlusion and seamlessly merging the individual slices to generate a precise and comprehensive representation of the scene geometry.

Given a hologram H of size $L \times L$, the Angular Spectrum Method (ASM) [9] is used to acquire a reconstruction volume by computing N numerical reconstructions at uniformly spaced distances z_i within a predefined depth interval of $[z_{\min}, z_{\max}]$, where $z_{\min}$ and $z_{\max}$ represent the minimum and maximum reachable depth values, respectively:

$$z_i = \frac{z_{\max} - z_{\min}}{N} \times i + z_{\min}. \tag{1}$$

The numerical reconstruction at a specific distance z_i is given as follows:

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

where f_x and f_y are the spatial frequencies along the X and Y direction of the hologram plane, λ is the acquisition wavelength, and z_i is the reconstruction depth. Only the amplitude of the numerical reconstructions is retained, constituting the reconstruction volume. The reconstruction volume is sliced horizontally as follows

$$M_y(i,x) = \left|\mathcal{P}_{z_i}\{H\}\right|(x,y) \quad i = 0, ..., N. \tag{3}$$

As explained in [7], each numerical reconstruction is sharp only at the part of the scene with a depth equal to the used reconstruction distances. Thus when slicing the reconstruction volume into horizontal slices, each slice M_y will be contaminated with speckle noise, with the exception of a few sharp lines. These lines correspond to the intersection points between the 3D scene and the horizontal plane with an elevation equal to y, as depicted in Figure 1. In order to accurately reconstruct the geometry of the scene, it is necessary to extract the in-focus points forming these sharp lines from each of the horizontal slices.

In the present work, a neural network is supervised to segment the in-focus regions from the input horizontal slice. The segmentation problem can be mathematically formalized as

$$\hat{I}_y = \mathcal{G}(M_y), \tag{4}$$

$$\mathcal{L} = BCE(\hat{I}_y, I_y), \tag{5}$$

where $\mathcal{G}$ is implemented using a residual U-net [10] architecture, $\hat{I}_y$ and I_y are the estimated and the ground truth in-focus maps, and $\mathcal{L}$ is the loss function, based on binary cross entropy function BCE.

The segmentation problem outlined in Eq. (4) is characterized by an imbalance, with only a limited number of in-focus regions on the horizontal slice. As a result, training the network $\mathcal{G}$ without an appropriate loss function can result in suboptimal model performance and biased predictions. To overcome this challenge, the segmentation loss must be adapted to make sure that the network converges. One strategy that has been shown to be effective is the use of a weighted sum of the boundary loss [11] and binary cross-entropy. This approach stabilizes the network training process and leads to a significant improvement in the final results. The weighted loss is given as

$$\mathcal{L} = \mathcal{L}_{BCE} + \alpha\mathcal{L}_B, \tag{6}$$

where $\mathcal{L}_{BCE}$ is the binary cross-entropy, $\mathcal{L}_B$ is the boundary loss [11], and α is a hyperparameter set to 0.5 in the experiments.

The combination of boundary loss and binary cross entropy loss used in this method will result in the generation of thick in-focus lines, where each point along the X-axis may have multiple associated depth values. To address this issue and extract a single, accurate depth value per point, we assume that the ground truth in-focus lines represent the skeleton of the predicted thick lines. Using this assumption, we can approximate the accurate depth value per point by taking the median along the Z-axis:

$$\hat{d}_y = median\{z : \hat{I}_y(z,x) > 0 \quad z = 0, ..., N-1\} \tag{7}$$

where $\hat{d}_y$ is the final predicted depth value with an elevation equal to y, and $\hat{d}$ is the final depth maps given as:

$$\hat{d}(y,x) = \hat{d}_y(x). \tag{8}$$

The proposed H-Seg and the patch-based approach [8] have been trained on the same dataset, which consists of 1400 holograms that are equally divided into three classes: *Piano*, *Table*, and *Woods*.

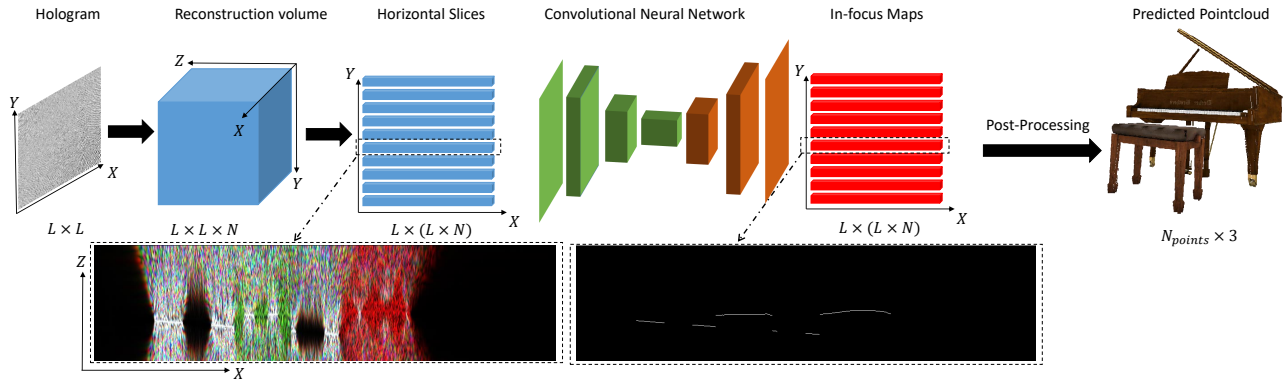


Figure 1: Illustration of the different steps of the proposed method.

These holograms were acquired using the layer-based methods described in [12] with a resolution of 1024 × 1024 and four different pixel pitches: 8, 6, 4, and 2μm. The evaluation phase included the use of two hologram sets. The test set comprised 300 holograms, created by adding 100 additional acquisition angles to each training scene (*Piano*, *Table*, *Woods*). The validation set consisted of 200 holograms recorded from two additional scenes (*Cars*, *Dices*) unseen during the training phase. For each hologram, a reconstruction volume was created by computing $N = 256$ numerical reconstructions at uniform intervals within the depth range $[z_{\min}, z_{\max}]$, where $z_{\min}$ and $z_{\max}$ are the minimum and maximum reachable depth values in the dataset. These values are manually computed from the dataset and set according to the used pixel pitch. Each numerical reconstruction took 1.31 second to complete using a Matlab implementation on 11th Gen Intel Core i9-11900F. Both the patch-based and H-Seg methods were trained for 200 epochs with a batch size of 256. An early stop was implemented, halting training if the test accuracy did not improve for 10 epochs. Additionally, an exponential learning decay was applied with a step size of 10 and a rate of 0.8 chosen through experimentation and hyperparameter tuning. For the patch-based method, at each iteration the images that make up the cropped reconstruction volume are randomly shuffled and then all images are subjected to the same random flip (horizontally or vertically with equal probability). In contrast, the proposed H-Seg method involves independently flipping each slice and then applying a random translation along the X and Z directions. The use of early stopping, learning decay, and data augmentation helped to improve the generalization performance of the networks, resulting in better performance on unseen data. The patch-based and H-Seg were evaluated not only for their ability to predict depth accurately but also for their ability to perform empty background removal. To assess this ability, an additional background class was added to the patch-based method and the networks of both methods were trained to classify any background points into this class. The performance of the tested methods was evaluated using the ℓ_1 norm between the predicted and ground truth depth maps. Table 1 gives the obtained results when the two networks are trained and evaluated using the same pixel pitch.

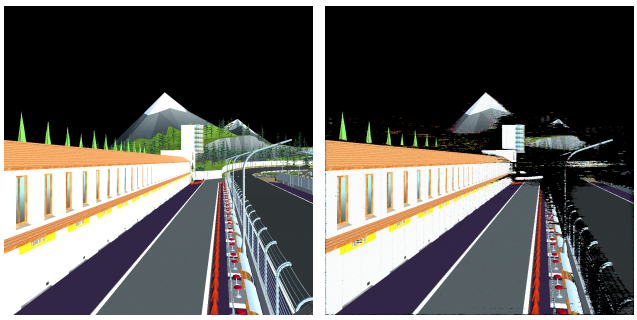
	Piano	Table	Woods	Cars	Dices
	Pixel pitch : 8μm				
Patch-based approach	2.93/79.09	4.36/106.17	2.39/ 84.09	6.92/ 41.19	5.22 / 12.22
Proposed H-Seg approach	0.14/0.61	0.32/1.14	0.25/0.97	2.02/7.35	1.03/3.96
	Pixel pitch : 6μm				
Patch-based approach	3.46/ 58.88	6.55/70.64	3.12/85.72	10.06/58.5	8.4/38.89
Proposed H-Seg approach	0.49/3.24	0.75/4.11	0.79/3.93	2.23/7.9	1.5/8.07
	Pixel pitch : 4μm				
Patch-based approach	5.02/ 71.52	7.2/73.62	3.23/76.26	12.55/64.06	12.65/69.23
Proposed H-Seg approach	0.69/4.44	0.97/5.18	1.24/4.9	3.72/11.96	2.33/10.43
	Pixel pitch : 2μm				
Patch-based approach	14.75/54.1	20.55/ 64.52	4.37/58.19	16.59/ 61.36	19.74/ 33.08
Proposed H-Seg approach	1.57/8.93	1.42/5.77	1.62/6.2	5.35/17.72	5.1/25.35

Table 1: The table presents the ℓ_1 norm distance between the ground truth and predicted depth maps in terms of reconstruction planes number. Two scenarios are analyzed, with the right term calculated using the entire depth map and the left term considering only foreground objects. The network is trained from scratch for each pixel pitch and evaluated on both test and validation sets.

Depth accuracy : Results show that the patch-based method exhibits a significant discrepancy between its accuracy for in-object pixels and its overall accuracy for the depth map, indicating that the network is unable to effectively distinguish between foreground and background pixels. In contrast, the proposed H-Seg method performs better at separating the background and maintaining a smaller prediction gap. However, this gap tends to widen for smaller pixel pitches.

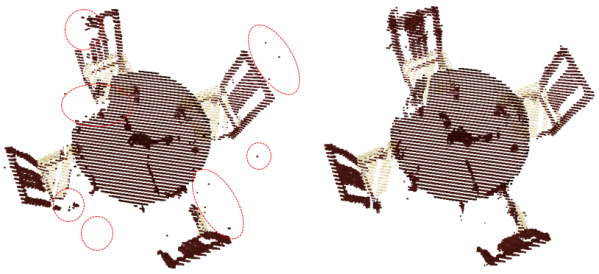
The difference in prediction accuracy between the two approaches can be attributed to their respective problem formulations. The proposed H-Seg method uses horizontal slices that span the entire scene via horizontal planes perpendicular to the hologram, providing global information about the scene. This allows for accurate prediction of background pixels. On the other hand, patch-based methods rely on local information from the crop reconstruction volume, ignoring the temporal correlation between the images that compose the volume. As a result, the network is unable to learn the relevant features that distinguish background pixels, leading to the observed gap in prediction.

Figure 2 displays the inferred in-focus map for a distinct scene that includes background elements. Remarkably, despite being trained on scenes with no back-



(a) (b)

Figure 2: (a) Ground-truth and (b) inferred in-focus map using the proposed method on a scene with a non-zero background.



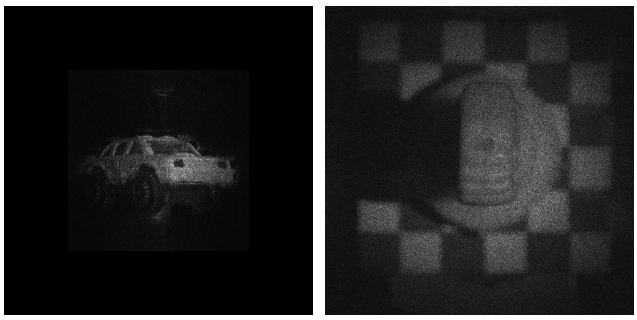
(a) (b)

Figure 3: Comparison of pointclouds predicted using the (a) patch-based and (b) H-Seg methods. The patch-based method produces a point cloud with significant scatter due to the poor performance along edges. As the proposed H-Seg method uses a horizontal slice that provides global information about the X, Z plane, it produces a smoother pointcloud with less scatter.

ground, the network accurately recovered the geometry of both scenes. However, some areas were still misclassified, indicating that there is room for further improvement in the model performance.

Generalization ability : Both methods show an increase in the ℓ_1 norm and prediction gap between the test and validation set as the pixel pitch becomes smaller. However, this gap is smaller when using the proposed H-Seg approach. This can be attributed to the data augmentation applied during network training, which artificially introduces variations and helps the network learn the underlying patterns in the data more effectively, leading to more accurate predictions on unseen data. While the patch-based method also utilizes data augmentation, it only introduces local variations to the reconstruction volume without changing the shape of the segmented region. In contrast, the proposed H-Seg method randomly flips and shifts each horizontal slice, and therefore the in-focus lines along the X or Z axis, to cover every possible region of the horizontal plane, leading to better performances.

In addition to synthetic scenes, we evaluated the proposed network on optically-acquired holograms provided by the Universidade da Beira Interior [13] shown in Figure 4, using 256 numerical reconstructions uniformly sampled between 1 to 256 mm. The obtained average error is 31.2mm for the Car scene and 28.1mm for the cube scene.



(a) Car (b) Horse

Figure 4: In-focus map inferred using the proposed method on optically acquired holograms [13].

Qualitative comparison of obtained depth maps : The patch-based method tends to produce depth maps with more roughness and discontinuities compared to the proposed H-Seg method, as demonstrated in Figure 3. There are two factors contributing to this issue. Firstly, a low-resolution patch size is used for network training and inference to avoid excessive GPU usage when dealing with large reconstruction intervals. Secondly, the poor performance at the edges of the patches, resulting from the convolution operation which extends beyond the patch boundary and thus produces unreliable or nonsensical results. In contrast, the proposed H-Seg method uses a horizontal section that fully spans the X and Z dimensions, resulting in fewer discontinuities and smoother depth maps.

Computational complexity : In addition to its accuracy, the proposed H-Seg method has the advantage of being faster (14.30 second per reconstruction volume) than the patch-based method (26.05 second per reconstruction volume). When processing a full reconstruction volume, H-Seg only requires L/b (4 in this experiment) inferences with a batch size of b, while the patch-based method necessitates $(L \times L)/(s \times s)$ (1024 in this experiment) inferences with a fixed batch size of N. As a result, H-Seg not only requires fewer inferences to process the reconstructions, but also has lower GPU requirements per batch that are not linearly correlated with the number of numerical reconstructions.

Overall, the proposed H-Seg approach yields faster and more accurate results compared to the patch-based method. However, the process of extracting horizontal slices requires the computation and storage of large amounts of numerical reconstructions, which can pose a challenge for high-resolution holograms. In future work, we plan to investigate interpolation techniques to reduce the number of numerical reconstructions required to retrieve the depth map of the scene. In addition, we plan to enhance the proposed method by incorporating temporal information between the horizontal slices that constitute the reconstruction volume, or by using cross slices in both X and Y directions with an additional merge and refine step.

References

[1] David Blinder, Ayyoub Ahar, Stijn Bettens, Tobias Birnbaum, Athanasia Symeonidou, Heidi Ottevaere, Colas Schretter, and Peter Schelkens. Signal processing challenges for digital holographic video display systems. *Signal Processing: Image Communication*, 70:114–130, 2019.

[2] Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation: A survey. *Computer Vision and Image Understanding*, 134:1–21, 2015. Image Understanding for Real-world Distributed Video Networks.

[3] Said Pertuz, Domenec Puig, and Miguel Angel Garcia. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46(5):1415–1432, 2013.

[4] Xin Fan, John J. Healy, and Bryan M. Hennelly. Investigation of sparsity metrics for autofocusing in digital holographic microscopy. *Optical Engineering*, 56(5):053112, 2017.

[5] Miu Tamamitsu, Yibo Zhang, Hongda Wang, Yichen Wu, and Aydogan 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 *BiOS*, 2017.

[6] Victor Dyomin and D. V. Kamenev. A comparison of methods for evaluating the location of the best focusing planes of particle images reconstructed from digital holograms. *Russian Physics Journal*, 56:822–830, 2013.

[7] Nabil Madali, Antonin Gilles, Patrick Gioia, and Luce Morin. Automatic depth map retrieval from digital holograms using a depth-from-focus approach. *Appl. Opt.*, 62(10):D77–D89, Apr 2023.

[8] Nabil Madali, Antonin Gilles, Patrick Gioia, and Luce Morin. Automatic depth map retrieval from digital holograms using a deep learning approach. *Opt. Express*, 31(3):4199–4215, Jan 2023.

[9] Joseph W Goodman. Introduction to fourier optics. *Introduction to Fourier optics, 3rd ed., by JW Goodman. Englewood, CO: Roberts & Co. Publishers, 2005*, 1, 2005.

[10] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[11] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In M. Jorge Cardoso, Aasa Feragen, Ben Glocker, Ender Konukoglu, Ipek Oguz, Gozde Unal, and Tom Vercauteren, editors, *Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pages 285–296. PMLR, 08–10 Jul 2019.

[12] Antonin Gilles, Patrick Gioia, Rémi Cozot, and Luce Morin. Hybrid approach for fast occlusion processing in computer-generated hologram calculation. *Appl. Opt.*, 55(20):5459–5470, Jul 2016.

[13] Marco V. Bernardo, Pedro Fernandes, Angelo Arrifano, Marc Antonini, Elsa Fonseca, Paulo T. Fiadeiro, António M.G. Pinheiro, and Manuela Pereira. Holographic representation: Hologram plane vs. object plane. *Signal Processing: Image Communication*, 68:193–206, 2018.