2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Binary Amplitude-Only Hologram Generation for Acoustic End-Effector Design by Physics-based deep learning

Qing Liu, Hu Su, Jiaqi Li, Youfu Li, Zhiyuan Zhang, and Song Liu, *Member, IEEE*

***Abstract*—Acoustic holography has emerged as a cutting-edge technique for constructing a micro-robot acoustic end-effector for non-contact manipulation. As one of typical implementations of acoustic holography, Binary Amplitude-Only Hologram (BAOH) featured with a simple structure provides an efficient alternative for modulating acoustic fields that support micro-robotic manipulation. In the present study, we propose a deep learning based BAOH generation method for constructing precise and high-resolution end-effector based on acoustic field. Specifically, we model the BAOH generation problem into an optimization framework. The framework combines an acoustic wave propagation model with the deep neural network, in favor of bypassing the laborious collection of labeled data and facilitating the model to learn the inverse mapping. Additionally, to address the issues of gradient invalidation and information loss caused by binarization, the framework uses an adaptive binarization layer consisting of differentiable binarization and adaptive threshold automatically learned during training, which facilitates to realize end-to-end optimization and increase the non-linear capacity of the model. The simulation experiments show that the proposed method is capable to predict BAOH that supports precise, robust, versatile and real-time construction of acoustic end-effector, enjoying broad prospects in various applications related to micro-robotic manipulation.**

I. Introduction

Over the past years, the rapid development of microscopic techniques and micro-scale manufacturing technology have greatly promoted micro-robotic manipulation in the realm of robotics. Conventional contact manipulation mechanism may encounter challenges of positioning accuracy constraints, damage risks and adhesion forces in the micro-scale domain [1]. In contrast, non-contact mechanism has delightful characteristics of non-invasiveness and high precision, providing itself with high applicability to micro-scale manipulation [2]. A range of physical field-based end-effectors have been developed for non-contact robotic manipulation. Within this category, acoustic field [3] presents unique advantages over its counterparts in optics [4] and magnetics [5], including high spatial resolution, low energy cost, sufficient penetration depth, non-injury, and reliable functionality in various working mediums. These advantages imprint acoustic driven manipulation with favorable bio-compatibility and high adaptability, making it well-suited for comprehensive applications such as drug-delivery [6], and in-vivo clinical surgery [7].

Over the past decades, acoustic holography has emerged as a revolutionary tool for non-contact manipulation with its flexibility of shaping the acoustic wavefront and modulating the acoustic field. The modulated field can be regarded as a non-invasive acoustic end-effector to maneuver the objects by the interaction between the objects and the acoustic waves, allowing the operations of trap [8], levitation [9], and rotation [10], etc. To date, there are two dominant approaches for the implementation of acoustic holography, encompassing phased array transducers (PAT) [11] and metasurface [12]. PATs use individually controllable wave-emitters to generate dynamically tunable and reconfigurable wavefronts via wave superposition. However, since every addressable transducer requires independent support driving circuit, such complex fabrication leads to limited controllable elements, insufficient resolution and tedious calibration process, which impose limitations on the degree of freedom and hinder more precise manipulation especially at the microscale [13]. In comparison, in the form of two-dimensional metamaterials of subwavelength thickness [12], metasurface provides a viable light-weight device to tailor the acoustic fields with diffraction-limited resolution at relatively low cost.

Among various kinds of acoustic holograms, binary amplitude-only hologram (BAOH), featured by encoding amplitude profile with binary digits, inherently aligns well with metasurface. Taking advantage of acoustic impedance mismatch principle, a border between two different mediums could block the transmission of incident waves, which enables a patterned metasurface encoding the BAOH to steer the acoustic beams. The simple structure of BAOH avoids the relatively sophisticated and intricate fabrication process that is indispensable for phase retardation based metasurface [12]. Moreover, BAOH is compatible with surface micro-machining technology and programmable metasurface [14] to fulfill spatial ultrasound modulation, showcasing its capability in dynamic acoustic robotic manipulation.

Computer-Generate holography (CGH) methods are proposed to digitally record and synthesize holographic interference patterns. The iterative ones, such as iterative angular spectrum approach (IASA) [15], iterative

* This work was supported in part by National Natural Science Foundation of China under Grand 62303321 and in part by Hong Kong Research Grant Council under CityU11206122. (Q. Liu and H. Su contribute equally to this work.) (Corresponding author: *Song Liu*).

Q. Liu, J. Li is with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China (Email: liuqing2022@shanghaitech.edu.cn; lijq1@shanghaitech.edu.cn).

H. Su is with the Institute of Automation, Chinese Academy of Science, Beijing, China (Email: hu.su@ia.ac.cn).

Y. Li is with the Department of Mechanical Engineering, City University of Hong Kong, Kowloon, HK (email: meyfli@cityu.edu.hk)

Z. Zhang is with the Department of Mechanical and Process Engineering, ETH Zurich, Zurich, Switzerland (e-mail: zhiyzhang@ethz.ch)

S. Liu is with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China, and with Shanghai Engineering Research Center of Intelligent Vision and Imaging, Shanghai, China (e-mail: liusong@shanghaitech.edu.cn).

backpropagation [3], weighted Gerchberg-Saxton [16], and Diff-PAT [17] suffer from high computation cost and suboptimal convergence, making them not suitable for robotic manipulation in practice that requires high efficiency and precision. Recently, the success of deep learning (DL) [18] in interdisciplinary applications has encouraged the society to delve into leveraging deep neural networks for acoustic hologram generation [19, 20]. Nevertheless, mainstream researches mentioned above predominantly focus on phase and/or amplitude holograms and the research regarding BAOH is relatively limited. In the context of BAOH, binarization of amplitude profile will inevitably invalidate the optimization process and discarding of phase profile will bring about information loss. Thus, there is a pressing need for mitigating the information loss and obtaining BAOH that supports acoustic manipulation with high-fidelity.

This paper proposes a novel physics-based deep learning framework for BAOH calculation. Incorporated with acoustic wave propagation model, the framework is realized in a self-supervised learning manner with no need of collecting paired data. To mitigate the information loss and tackle gradient invalidation caused by traditional binarization, the framework adopts an adaptive binarization layer that includes a differentiable binarization function and automatically learnt adaptive threshold. The bespoke layer is embedded in the training process to realize end-to-end learning. The experiments showcase that the proposed framework provides a feasible and effective alternative for robotic manipulation that surpasses existing BAOH calculation algorithms in aspects of both precision and real-time performance.

II. Acoustic Robotic Manipulation based on BAOH

A. Manipulation with Acoustic Micro-Robot End-Effector

Arbitrarily defined and user-specific acoustic fields have sprung up as a cutting-edge end-effector for non-contact micro-robot manipulation, where tailored wavefronts can interact with the operational target and implement robotic manipulation techniques such as acoustic levitation [21] and acoustic tweezers [3]. The mechanism behind manipulation with acoustic fields primarily involves the impact of sound waves on the object through the exertion of acoustic radiation force (ARF), which enables acoustic field to manipulate objects whose densities or compressibilities differ from those of their surrounding media. Therefore, acoustic field configured into unique morphology could serve as a non-contact end-effector in support of customized micro-robotic manipulation. Frequently-used morphologies include single foci, twin trap, and vortex [3]. Besides, acoustic fields encoding with arbitrary patterns permit continuous migration of particles along trajectories [13]. These accomplishments showcase the flexibility of acoustic fields in dynamic and reconfigurable manipulation.

Fig. 1 illustrates the encoding and reconstructing the acoustic field-based end-effector implemented with BAOH. By using BAOH, the morphology of the acoustic field at the target plane can be tuned by modifying sound transmission of incident waves. The transmitted acoustic waves, which encodes with binary amplitude profile and uniform phase profile, propagate through the working medium and form an interference field at the target plane following the superposition principle. Given a requested acoustic manipulation procedure, encoding the source BAOH with corresponding acoustic field can be articulated as an ill-posed inverse problem. It is a key concern to seek for the optimal BAOH generation method with the goal to achieve satisfactory reconstruction performance and thus fulfill precise robotic manipulation.

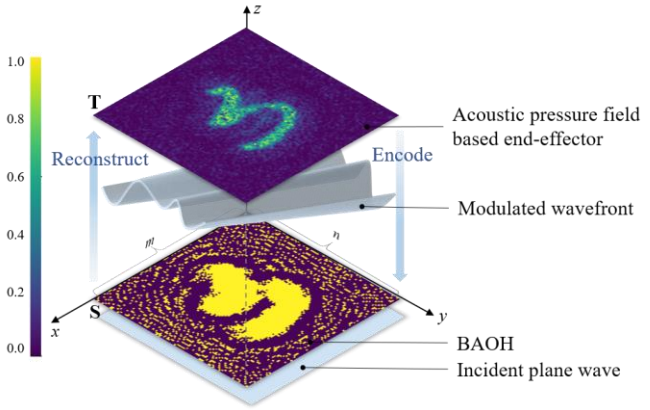

Figure 1. Schematic of acoustic micro-robotic end-effector implemented with BAOH.

B. Acoustic Field Generation with BAOH

To mathematically formulate the inverse problem, we establish the definitions as follows. The source plane is denoted as **S**, and the target plane is denoted as **T**, which are both discretized into $m \times n$ pixels. The forward and backward acoustic wave propagation between **S** and **T**, i.e. reconstruction and encoding, are two symmetric processes modelled by the angular spectrum method (ASM) [22]. In the setting of BAOH, the source amplitude profile A_S is encoded with binary digits and the phase profile ϕ_S is uniformly distributed, denoted as ϕ_U. The forward propagation f from the source BAOH to target pressure field is formulated as:

$$f:(A_S,\phi_U)_{\mathbf{S}} \xrightarrow{\text{ASM}} (p_C)_{\mathbf{T}} \quad s.t.\ \phi_U \sim U(0); A_S \in \{0,1\}^{m\times n} \tag{1}$$

where $U(0)$ denotes the uniform distribution with a fixed value of 0, $A_S \in \{0,1\}^{m\times n}$ consists with $m \times n$ pixels that have the values of either 0 or 1, and p_C represents the reconstructed acoustic pressure field on **T**. Intuitively, given the expected pressure field p_E, the inverse problem f^{-1} can be articulated as:

$$f^{-1}:(p_E)_{\mathbf{T}} \xrightarrow{\text{ASM}} (A_S)_{\mathbf{S}} \quad s.t.\ A_S \in \{0,1\}^{m\times n} \tag{2}$$

The encoding process formulated in (2) is nonlinear, non-convex, and may have multiple solutions or even fail to converge. Thus, we formulated the BAOH generation problem into an optimization process, the aim of which is to minimize the difference between p_E and p_C reconstructed by the source BAOH, which can be articulated as:

$$\arg\min_{A_S}[diff(Norm(p_E), Norm(f(f^{-1}(p_E),\phi_U))] \tag{3}$$

where $Norm(.)$ represents the normalization operation that restrains the acoustic pressure into range [0, 1].

III. PHYSICS-BASED DEEP LEARNING FOR BAOH GENERATION

As a data-driven method, training DL models to fit the inverse mapping problems requires extensive labeled data pairs, which is a challenge within the scope of BAOH. Specifically, collecting the source holograms and corresponding reconstructed acoustic fields with the hydrophone could be extremely time-consuming and labor-intensive. Moreover, the domain bias between the ideal acoustic fields and data captured from realistic scenarios may lead to unstable predictions, degraded performance or even model failure. To address the issues, we propose a self-supervised learning framework that combines a physical model and deep learning. Additionally, we introduce a novel adaptive binarization layer to tackle the gradient invalidation and information loss ascribed to the binarizing procedure of BAOH. Further details are provided in this section.

A. Physics-based Deep Learning Framework

As showed in Fig. 2, the framework for predicting the optimal BAOH comprises a multi-layer convolutional neural network (CNN), an adaptive binarization layer, and an acoustic wave propagation model. The entire framework is trained end-to-end while in test phase, only multi-layer CNN is used followed by a post binarization to deduce final BAOH. We utilize U-net [23] as the CNN architecture, which is featured with an encoder-decoder structure and frequently used in reconstruction applications. The depth of the U-net is determined by the size of data fed into the CNN, in purpose of achieving a sufficient receptive field capable to extract global features. In the current design, both encoder and decoder have six blocks. Each encoder block consists of two conv layers, including a 3x3 convolutions, Batch Normalization (BN) and rectified linear unit (ReLU), followed by a 2x2 max pooling operation. Correspondingly, each decoder block consists of an upsampling of the feature map followed by a 2x2 upsampling convolution ("Up-Conv"), a concatenation with the corresponding feature map from the contracting path, and two 3x3 convolutions, each followed by BN and ReLU. At the final layer, 1x1 convolution is used to predict the probability map and corresponding threshold map.

The self-supervised learning is conducted with the assistance of physical model. To be specific, p_E is fed into the U-net to predict both the amplitude-only hologram (AOH) and the adaptive threshold map, which are then combined to calculate the approximate BAOH through a differentiable binarization function. The subsequent wave propagation modelling with ASM is embedded in the framework to promote the CNN to learn the intrinsic physical characteristics of the inverse mapping problem. With the physical model, the reconstruction results are obtained through the approximate BAOH as described in (1). The CNN is directly penalized mainly on the differences between the reconstructed and expected fields during training. In the inference period, we perform a hard binarization with the learnt threshold map on the AOH to guarantee every pixel of the obtained standard BAOH have the value of 0 or 1. The standard BAOH is leveraged for the subsequent practical application of acoustic end-effector.

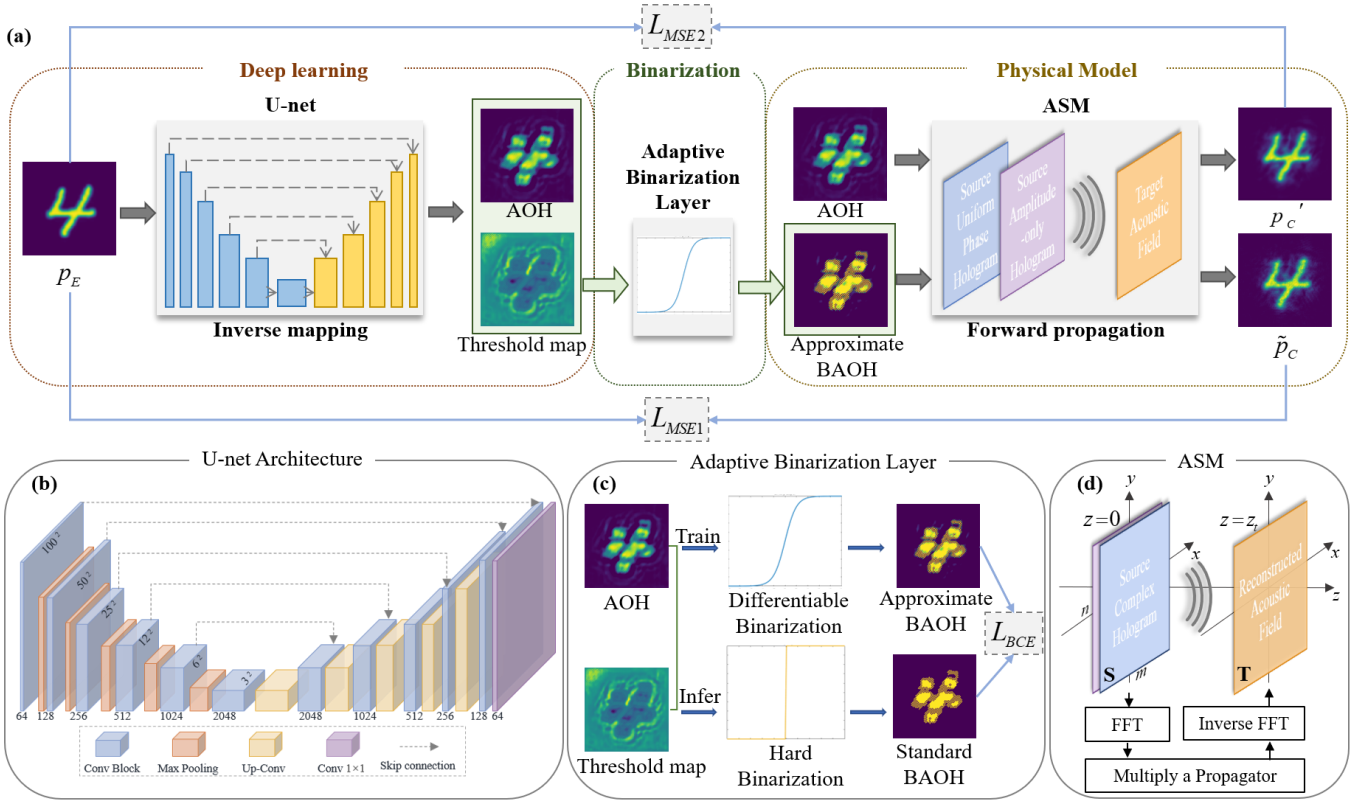

Figure 2. Illustration of physics-based deep learning framework. (a) illustrates the training of the framework, while the key components are detailed in (b), (c), and (d). (b) shows the U-net architecture. (c) describes the schematic of adaptive binarization layer. (d) is the flowchart of forward wave propagation modelling with the ASM method.

B. Adaptive Binarization Layer

With regard to BAOH, the demand of binary amplitude profile will lead to ineffective optimization in CNN due to the gradient invalidation caused by the non-differentiable binarization process. Additionally, the selection of threshold of binarization holds great importance, as it directly determines the quality of binarized output. However, binarization with a predefined fixed threshold can be crude and arbitrary, which consequently degrade the reconstruction quality. In this section, we propose a novel adaptive binarization layer to address the above issues. In the layer, differential binarization function is presented to preserve the gradient flow during training and meanwhile, pixel-wise threshold map is learned for binarization. The proposed adaptive binarization layer is detailed as follows.

Given the expected acoustic field p_E, the CNN will produce the corresponding prediction of AOH and the pixel-wise threshold map. For clarity, AOH is denoted as $A \in \mathbb{R}^{m\times n}$, and threshold map as $T \in \mathbb{R}^{m\times n}$, where m and n indicate the width and height of the map respectively. The AOH is normalized to the range of [0,1], while the threshold map is under no constraints. To preserve the gradient flow during and retain the information content to some extent, we apply the sigmoid function to approximate the non-differentiable binarization function. In the training phase, the approximate BAOH is yielded by the differentiable binarization formulated as:

$$\tilde{B}_{(m,n)} = \frac{1}{1+e^{-k(A_{(m,n)}-T_{(m,n)})}} \tag{4}$$

where $\tilde{B}$ presents the approximate BAOH, k is a hyper-parameter that describes the fitting degree, which is set to 1 empirically in the study, and (m,n) describes the coordinate of every pixel in the map.

In the inference phase, the standard BAOH is yielded by conducting the hard binarization by comparing AOH and threshold map pixel by pixel, which is formulated as:

$$B_{(m,n)} = \begin{cases} 1, \ if \ A_{(m,n)} > T_{(m,n)} \\ 0, \ otherwise \end{cases} \tag{5}$$

where B presents the standard BAOH. By incorporating the adaptive binarization layer, the entire framework could be trained end-to-end. Besides, adaptive threshold map alleviates the information loss caused by binarization, thus improving the reconstruction quality of BAOH.

C. Loss Function

In pursuit of the optimal reconstruction performance of BAOH, the CNN is primarily penalized on the disparity of the expected pressure field p_E and the approximate BAOH-reconstructing pressure $\tilde{p}_C$, quantified by the pixel-wise mean squared error (MSE). The MSE loss denoted as L_{MSE1} is formulated as

$$L_{MSE1} = \frac{1}{mn}\sum (p_E - \tilde{p}_C)^2 \tag{6}$$

Since the AOH theoretically determines the upper limit of the reconstruction quality of BAOH, the enhancement of the reconstruction quality of AOH is also integrated into the training process of CNN. Specifically, we apply the MSE loss likewise, to measure the difference between p_E and $p_C^{`}$ reconstructed from AOH. The loss is denoted as L_{MSE2}, and formulated as

$$L_{MSE2} = \frac{1}{mn}\sum (p_E - p_C')^2 \tag{7}$$

To minimize the gap between the differentiable binarization and hard binarization, we additionally utilize a binary cross-entropy (BCE) loss to supervise the approximate BAOH, with standard BAOH as labels. The BCE loss is denoted as L_{BCE}, and formulated as

$$L_{BCE} = \frac{1}{mn}\sum B\log(\tilde{B}) + (1-B)\log(1-\tilde{B}) \tag{8}$$

The loss function L can be expressed as a weighted sum of the aforementioned losses:

$$L = L_{MSE1} + \alpha L_{MSE2} + \beta L_{BCE} \tag{9}$$

where α and β are set as 0.01 and 0.05 respectively, according to the experimental goal and the numerical values of the losses.

IV. Experiments and Results

Comprehensive experiments were conducted to evaluate the viability and superiority of our proposed framework. The section firstly introduces the implementation details and evaluation metrics. Afterwards, ablation study on the losses, feasibility assessment and comparison with state-of-the-art (SOTA) are provided to demonstrate the effectiveness of the proposed method.

A. Implementation Details and Evaluation Metrics

Our method was implemented by using PyTorch on an Ubuntu 18.04 server equipped with 4 Nvidia Tesla M40 GPUs and an Intel Core i7- 10700 CPU with a frequency of 2.90G Hz. The experiments were conducted on the dataset which includes 2000 images selected randomly from MNIST dataset [24]. The images were resized to 100×100 and normalized to range [0,1], in alliance with BAOH setting. The preprocessed images were stored in the dataset $\{p_E\}$ and randomly divided into training, validation, and test sets by the ratio of 8:1:1. The U-net was initialized by kaiming method and optimized by Adam optimizer. More detailed parameters for constructing the physical model and training the CNN are listed in Table I.

TABLE I. Parameters Used for Constructing the Physical Model and Training the CNN

Attribute	Parameter	Value
Physical Model	Sound speed	1480 m/s
	Excitation frequency	2.32 MHz
	Propagation distance	20 mm
	Hologram resolution	100 × 100 pixels
	Pixel size	320 μm
CNN	Learning rate	0.001
	Training epoch	500
	Batch size	16

To quantitatively evaluate the reconstruction performance, we employed several commonly-used evaluation metrics, including MSE, Peak Signal-to-Noise Ratio (PSNR), Structure Similarity Index Measure (SSIM) [25] and Efficacy [26]. MSE, PSNR, and SSIM aim at assessing the similarity between expected pressure field and reconstructed pressure field. Efficacy is applied to gauge the reconstruction efficiency, which is quantified by the overlap between expected and reconstructed pressure field. Among them, PSNR was used as a major metric in performance evaluation.

B. Ablation Study of Loss Function

The ablation study was conducted to analyze the effect of each component in the loss function (9). Table II shows the PSNR values of the reconstruction results under different loss combinations. The results indicate that every component of the loss function facilitates to enhance the reconstruction performance, and combining all three sub-losses achieves the highest PSNR. The effectiveness of the loss function in (9) is thus demonstrated.

TABLE II. ABALATION STUDY OF LOSS FUNCTION

L_{MSE1}	L_{MSE2}	L_{BCE}	PSNR
√			14.80 dB
√	√		15.08 dB
√		√	21.13 dB
√	**√**	**√**	**22.27 dB**

C. Reconstruction Performance Analysis

To showcase the efficacy of the training progress, the model was continuously evaluated on the validation set every 10 epochs during training. Fig. 3 provides the curves of training loss as well as the PSNR values, accompanied by perceptual reconstruction examples at three specific epochs. The curves indicate that the training loss decreased quickly during first epochs and then gradually converges to a low constant value, which is consistent with the evolutions of PSNR. The visualized examples suggest that as training proceeds, more energy concentrated on the foreground and less speckle noise existed in the background, contributing to an enhancing reconstruction quality.

To demonstrate the generalization of the proposed method, we tested the well-trained model on versatile images with distinct morphologies, which were randomly selected from MNIST, EMNIST [27], and Fashion MNIST [28], respectively. As shown in Fig. 4, the reconstruction results with satisfactory accuracy demonstrate the generalization ability and flexibility of the proposed method accordingly. The results further emphasize the potential of our method for constructing the acoustic end-effector applied across diverse scenarios.

D. Comparison with SOTA

For better illustration, comparison experiments were conducted with SOTA methods including IASA and Diff-PAT. Originally, IASA and Diff-PAT are proposed for phase-only holograms (POH) generation. Refer to [14], POH is converted into BAOH by setting the amplitudes to one where original phases lie in the range of $\pm\pi/2$, while remaining amplitudes are set to zero. In this manner, we obtained the BAOH through the POH derived by IASA and Diff-PAT. The BAOH generation methods are denoted as customized IASA and customized Diff-PAT to distinguish them from the original methods. The quantitative comparison results with the customized IASA and customized Diff-PAT are shown in the Fig. 5 and average numerical evaluation results are provided in Table II. We can see from the results that the proposed method outperforms the two SOTA methods in a large margin.

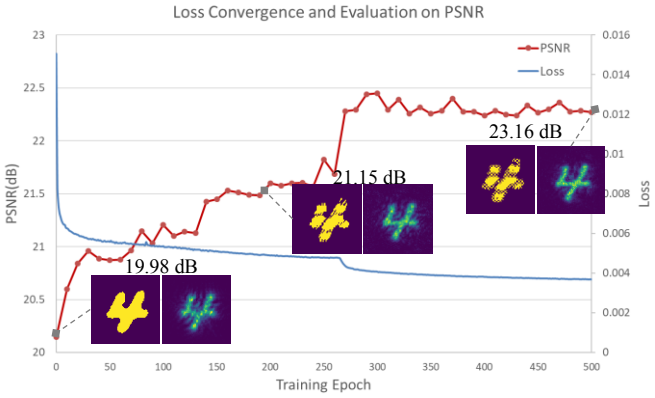

Figure 3. The training loss convergence and the average PSNR curve. One of the reconstructed acoustic pressure fields is visualized at the specific epochs of 1, 200, 500. The left pattern represents the predicted standard BAOH and the right pattern represents the corresponding reconstructed pressure field.

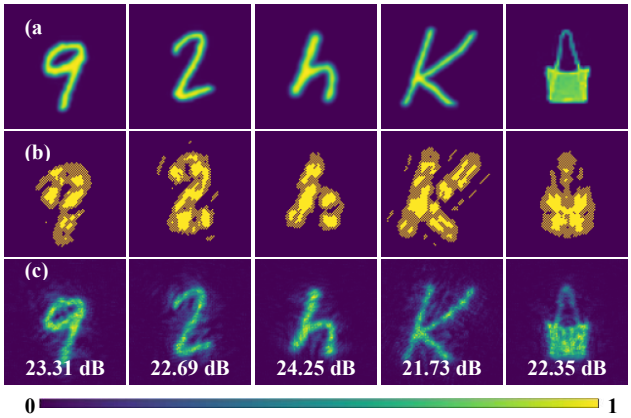

Figure 4. Generalization evaluation of the proposed model. (a) are the input expected acoustic pressure fields p_E. (b) are the predicted BAOH. (c) are corresponding reconstruction pressure fields p_C. The PSNR values are attached at the bottom of each column measuring the similarity of p_E and p_C.

Real-time performance is an essential concern in robotic manipulation. Table III compares the methods in terms of real-time performance. Our method is demonstrated to be about one order of magnitude faster than iterative methods. Note that the iterative optimization methods commonly require more iterations if input acoustic field with more sophisticated morphology or larger scale, while the CNN is less susceptible. The excellent real-time performance of our method holds the potential for future implementations such as programmable metasurface, providing a feasible alternative for dynamic and reconfigurable robotic manipulation.

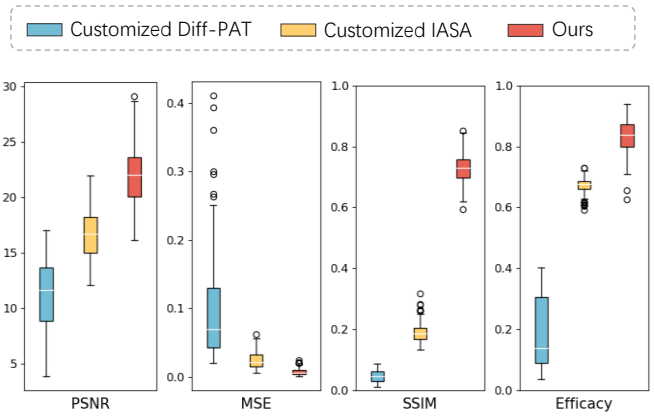

Figure 5. Box-and-whisker plot comparing proposed method with SOTAs regarding PSNR, MSE, SSIM and Efficacy.

TABLE III. QUANTITATIVE COMPARISON AMONG CUSTOMIZED SOTA METHODS AND OUR PROPOSED METHOD

Method	PSNR (dB)	MSE	SSIM	Efficacy	Computation time (sec)
Customized Diff-PAT	11.26	0.095	0.046	0.189	5.221
Customized IASA	16.66	0.024	0.189	0.673	0.176
Ours	**22.08**	**0.007**	**0.730**	**0.833**	**0.017**

V. CONCLUSION

This paper proposes a novel BAOH generation method for constructing acoustic field-based end-effector in support of micro-robotic manipulation. The innovative framework is composed with U-net, adaptive binarization layer, and wave propagation model formulated in ASM. The incorporation of physical model enables the U-net training in a self-supervised manner. The differentiable binarization process preserves the gradient flow during U-net training and the adaptive pixel-wise threshold map alleviates the information loss effectively. Ultimately, comprehensive experiments were conducted to verify the availability, versatility and robustness of proposed framework. In the future, the authors will delve into further enhancing the reconstruction performance and combining the BAOH-based end-effector with robotic arm to realize dexterous and flexible manipulation in real-world applications.

REFERENCES

[1] M. Savia and H. N. Koivo, "Contact Micromanipulation—Survey of Strategies," in *IEEE/ASME Transactions on Mechatronics*, vol. 14, no. 4, pp. 504-514, 2009.

[2] X. Liu *et al*., "Non-contact transportation and rotation of micro objects by vibrating glass needle circularly under water," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, Singapore, 2017, pp. 5996-6001.

[3] A. Marzo and B. W. Drinkwater, "Holographic acoustic tweezers," *Proceedings of the National Academy of Sciences* (*PNAS*), vol. 116, no. 1, pp. 84-89, 2019.

[4] A. Ashkin, and J. M. Dziedzic, "Optical trapping and manipulation of viruses and bacteria," *Science*, vol. 235, no. 4795, pp. 1517-1520, 1987.

[5] I. D. Vlaminck, and C. Dekker, "Recent Advances in magnetic tweezers," *Annu. Rev. Biophys*., vol. 41, pp. 453-472, Jan. 2012.

[6] J. J. Choi, R. C. Carlisle, C. Coviello, L. Seymour, and C. Coussios, "Non-invasive and real-time passive acoustic mapping of ultrasound-mediated drug delivery," *Phys. Med. Biol.*, vol. 59, no.17, pp.4861, 2014.

[7] M. A. Ghanem, A. D. Maxwell, Y. Wang, and M. R. Bailey, "Noninvasive acoustic manipulation of objects in a living body." *Proceedings of the National Academy of Sciences* (*PNAS*), vol. 117, no. 29, pp. 16848-16855, 2020.

[8] S. C. Takatori, R. D. Dier, J. Vermant, and J. F. Brady, "Acoustic trapping of active matter," *Nat. Commun.*, vol. 7, no. 1, pp. 10694, 2016.

[9] E. H. Brandt, "Suspended by sound," *Nature*, vol. 413, no. 6855, pp. 474-475, 2001.

[10] A. Anhäuser, R. Wunenburger, and E. Brasselet, "Acoustic rotational manipulation using orbital angular momentum transfer," *Phys. Rev. Lett.*, vol. 109, no.3, pp.034301, 2012.

[11] A. Marzo, S. A. Seah, B. W. Drinkwater, D. R. Sahoo, B. Long, and S. Subramanian, "Holographic acoustic elements for manipulation of levitated objects," *Nat. Commun.*, vol. 6, no. 1, pp. 1-7, 2015.

[12] B. Assouar, B. Liang, Y. Wu, Y. Li, J.C. Cheng, and Y. Jing, "Acoustic metasurfaces," *Nat. Rev. Mater.*, vol. 3, no. 12, pp. 460-472, 2018.

[13] K. Kolesnik, M. Xu, P. Lee, V. Rajagopal, and D. J. Collins, "Unconventional acoustic approaches for localized and designed micromanipulation," *Lab Chip*, vol. 21, no.15, pp. 2837-2856, 2021.

[14] Z. Ma, K. Melde, A. G. Athanassiadis, M. Schau, H. Richter, T. Qiu, and P. Fisher, "Spatial ultrasound modulation by digitally controlling microbubble arrays," *Nat. Commun.*, vol. 11, no. 1, pp. 4537, 2020.

[15] K. Melde, A. G. Mark, T. Qiu, and P. Fischer, "Holograms for acoustics," *Nature*, vol. 537, no. 7621, pp. 518-522, 2016.

[16] J. Zhang, Y. Yang; B. Zhu, X. Li, J. Jin, Z. Chen, Y. Chen, and Q. Zhou, "Multifocal point beam forming by a single ultrasonic transducer with 3D printed holograms," *Appl. Phys. Lett.*, vol. 113, no. 24, pp. 243502, 2018.

[17] T. Fushimi, K. Yamamoto, and Y. Ochiai, "Acoustic hologram optimisation using automatic differentiation", *Sci. Rep.*, vol.11, no.1, pp.12678, 2021.

[18] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[19] M. H. Lee, H. M. Lew, S. Youn, T. Kim, and J. Y. Hwang, "Deep Learning-Based Framework for Fast and Accurate Acoustic Hologram Generation," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control,* vol. 69, no. 12, pp. 3353-3366, 2022.

[20] C. Zhong, J. Li, Z. Sun, T. Li, Y. Guo, D. C. Jeong, H. Su, and S. Liu, "Real-Time Acoustic Holography with Physics-Based Deep Learning for Robotic Manipulation," *IEEE Trans. Autom. Sci. Eng.*, 2023.

[21] M.A.B. Andrade, N. Pérez, J.C. Adamowski, "Review of Progress in Acoustic Levitation," *Brazilian J. Phys.*, vol. 48, pp. 190-213, 2018.

[22] K. Matsushima and T. Shimobaba, "Band-limited angular spectrum method for numerical simulation of free-space propagation in far and near fields," *Opt. Express*, vol. 17, no. 22, pp. 19662-19673, 2009.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234-241.

[24] L. Deng, "The mnist database of handwritten digit images for machine learning research," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[26] M. H. Eybposh, N. W. Caira, P. Chakravarthula, M. Atisa, and N. C. Pégard, "High-speed computer-generated holography using convolutional neural networks," *Optics and the Brain*, pp. BTu2C-2, 2020.

[27] G. Cohen, S. Afshar, J. Tapson and V. Schaik, "EMNIST : an extention of MNIST to handwritten letters", *arXiv preprint arXiv*:1702.05373, 2017

[28] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv*:1708.07747, 2017.