Neural Architecture Search Generated Phase Retrieval Net for Real-Time Off-Axis Quantitative Phase Imaging

Xin Shu, Mengxuan Niu, Yi Zhang, Wei Luo, and Renjie Zhou

***Abstract*—In off-axis Quantitative Phase Imaging (QPI), artificial neural networks have been recently applied for phase retrieval with aberration compensation and phase unwrapping. However, the involved neural network architectures are largely unoptimized and inefficient with low inference speed, which hinders the realization of real-time imaging. Here, we propose a Neural Architecture Search (NAS) generated Phase Retrieval Net (NAS-PRNet) for accurate and fast phase retrieval. NAS-PRNet is an encoder-decoder style neural network, automatically found from a large neural network architecture search space through NAS. By modifying the differentiable NAS scheme from SparseMask, we learn the optimized skip connections through gradient descent. Specifically, we implement MobileNet-v2 as the encoder and define a synthesized loss that incorporates phase reconstruction loss and network sparsity loss. NAS-PRNet has achieved high-fidelity phase retrieval by achieving a peak Signal-to-Noise Ratio (PSNR) of 36.7 dB and a Structural SIMilarity (SSIM) of 86.6% as tested on interferograms of biological cells. Notably, NAS-PRNet achieves phase retrieval in only 31 ms, representing 15× speedup over the most recent Mamba-UNet with only a slightly lower phase retrieval accuracy.**

***Index Terms*—Phase retrieval, neural architecture search, quantitative phase imaging, real-time reconstruction.**

I. Introduction

QUANTITATIVE Phase Imaging (QPI) has been widely applied to biomedical imaging and material metrology. Among all approaches off-axis interferometry-based QPI methods (off-axis QPI) can offer high speed phase imaging due to their single-shot image acquisition feature [1], [2], [3]. However, retrieving the phase map from a recorded fringe pattern or interferogram requires several image processing steps, which means the specimens are usually not observed in real-time. In conventional approaches, the steps include (i) retrieving the wrapped phase (valued between $-\pi$ to π) from the object interferogram (e.g., the Fourier transform-based method [4]); (ii) calibrating the phase or compensating the phase aberration by using an additional interferogram captured in a sample-free region [5]; and (iii) unwrapping the phase (e.g., the Goldstein algorithm [6]). Among these steps, phase unwrapping is the most time-consuming. To expedite phase retrieval for real-time phase imaging, parallel computation using sophisticated Graphics Processing Units (GPUs) or Field Programmable Gate Arrays (FPGAs) has been implemented to accelerate phase unwrapping [7], [8]. However, specialized programming is required that hinders its generalization. Furthermore, obtaining the calibration interferogram for aberration compensation becomes particularly challenging when imaging dense samples, as a sample-free region may not be readily available.

In recent years, artificial neural networks (ANNs), including the widely used, U-Net have achieved simultaneous phase retrieval and elimination of the need for aberration compensation [9], [10], [11], [12]. Despite simplifying the imaging operation in off-axis QPI, further applying these ANNs for real-time phase imaging is potentially limited by the relatively large computational latency. It is known that the network inference accuracy and latency heavily depend on the network's architecture. Therefore, we need a strategy to identify an optimal network architecture that minimizes computational latency and maintains high accuracy for phase retrieval.

Neural architecture search (NAS) [13] is a technique to automatically find an optimal network architecture from a large architecture search space. NAS-generated networks have outperformed manually designed networks in many tasks, including classification [14], semantic segmentation [15], etc. SparseMask [16] is an end-to-end semantic segmentation NAS scheme. SparseMask has a network architecture search space that covers different skip connection strategies from the encoder to the decoder, which enables searching for optimal ways to fuse low-level features rich in spatial details and high-level features containing semantic information. In SparseMask, a differentiable NAS search strategy is used to relax the architecture search space from discrete to continuous, and the gradient descent is used to efficiently search for optimal skip connections. Taking both segmentation accuracy and connectivity sparsity into account, SparseMask attains

Received 20 December 2024; revised 13 June 2025; accepted 13 June 2025. Date of publication 18 June 2025; date of current version 27 June 2025. This work was supported in part by Hong Kong Innovation and Technology Fund under Grant ITS/148/20, Grant ITS/178/20FP, and Grant ITS/229/23FP; in part by the Croucher Foundation under Grant CM/CT/CF/CIA/0688/19ay; in part by the Research Grant Council of Hong Kong SAR under Grant 14209521; and in part by the National Natural Science Foundation of China/Research Grants Council of Hong Kong Joint Research Scheme under Grant N_CUHK431/23. *(Corresponding author: Renjie Zhou.)*

Xin Shu and Renjie Zhou are with the Department of Biomedical Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China (e-mail: shuxin@link.cuhk.edu.hk; rjzhou@cuhk.edu.hk).

Mengxuan Niu is with the School of Mechanical and Electrical Engineering, Shenzhen Polytechnic University, Shenzhen, Guangdong 518055, China (e-mail: mengxuanniu@szpu.edu.cn).

Yi Zhang is with the Institute of Data Science, The University of Hong Kong, Hong Kong Island, Hong Kong SAR, China (e-mail: yizhang101@connect.hku.hk).

Wei Luo is with the College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou 310027, China (e-mail: luo.wei@zju.edu.cn).

Digital Object Identifier 10.1109/LPT.2025.3581063

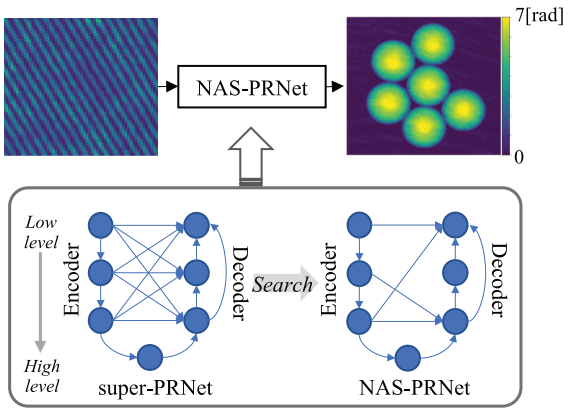

Fig. 1. Illustration of searching NAS-PRNet from a neural architecture search space through constructing the super-PRNet. Blue circles represent the encoder and decoder stages, lines with arrows represent the connections from head stages to tail stages.

a similar mean Intersection-over-Union (mIoU) but executes over three times faster, when compared with the widely-used Pyramid Scene Parsing Network (PSPNet) on the semantic segmentation PASCAL VOC 2012 test dataset.

To attain high-fidelity real-time phase imaging in off-axis QPI, we propose NAS-generated Phase Retrieval Net (NAS-PRNet) as illustrated in Fig. 1. The architecture of NAS-PRNet is derived through a customized adaptation of the SparseMask NAS algorithm. We then evaluate NAS-PRNet's performance in phase retrieval and compare it with the classic U-Net [17], the most recent Mamba-UNet [18] and EMCAD [19], and the original SparseMask to demonstrate its high accuracy and efficiency. Source code will be available at https://github.com/shuxin626/NAS-Phase-Retrieval-Net.

II. Methodology

To obtain the architecture of NAS-PRNet, an intermediate super-network for phase retrieval (super-PRNet), containing all possible encoder and decoder connections, is first constructed. Each connection in super-PRNet has an uniformly initialized weight of 0.5. Then, the super-PRNet is trained, and its encoder and decoder connections are pruned, according to the trained connection weights. After that, the architecture of NAS-PRNet is obtained. Finally, the NAS-PRNet is trained, and its performance is evaluated for phase retrieval. Both super-PRNet and NAS-PRNet are trained and tested with the same interferogram dataset. Here, we use biological cells to construct a dataset with diverse features and sizes for generalization. To enlarge the search space to include more possible network architectures, we modified the original SparseMask by loosening its constraints in two aspects: (1) allowing encoder features to propagate into all stages of the decoder, instead of only the corresponding lower-level decoder stages; and (2) applying a global sparsity loss to minimize the total number of connections, instead of a sparsity restriction for each decoder stage on a fixed quantity of connections. Furthermore, to fit SparseMask to the phase retrieval task, we modify its output layer (i.e., adopting a regression head), feature fusing style, kernel size in convolution, and feature depth strategy.

A. Structure of Super-PRNet

The structure of super-PRNet is illustrated in Fig. 2, where MobileNet-v2 is implemented as the encoder to efficiently

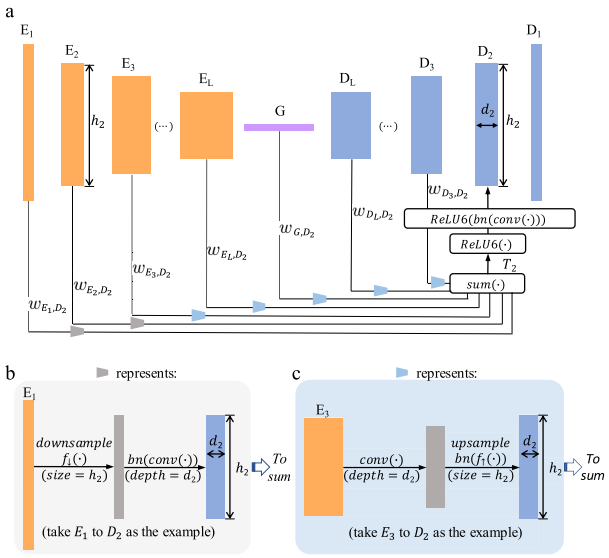

Fig. 2. Feature concatenation in super-PRNet. (a) Formation of D_2. E_i: i_{th} encoder feature; D_j: j_{th} decoder feature. (b) Processing of input features with spatial sizes larger than h_2. (c) Processing of input features with spatial sizes smaller than h_2.

extract multi-level features from an input interferogram. In Fig. 2(a), the encoder features at multiple levels are denoted as E_l, where l is the stage index ranging from 1 to $L = 8$ (i.e., 8 encoder features are input into the decoder). In addition, a ground encoder feature G is obtained by applying an average pool with a target size of 3×3, and then input into the decoder. The decoder integrates all possible connections between encoder and decoder stages, as well as different stages inside the decoder. Being symmetric with the encoder, the number of stages in the decoder is also L. D_l denotes the l_{th} decoder stage feature that has the same feature spatial size of h_l as E_l. The feature depth of D_l is set as $min(256, 8l)$. For the l_{th} encoder stage, its input features are $\{E_i | 1 \le i \le L\}$, G, and $\{D_j | l < j \le L\}$. As these input features differ in size and depth, we will efficiently fuse them in the following way: (i) for features $m \in M_l^+ = \{E_i | 1 \le i < l\}$ with spatial size larger than h_l, a bilinear down-sampling with a target spatial size of h_l is first applied before a convolution operation with a target feature depth of d_l; (ii) for features $m \in M_l^- = \{E_i | l \le i \le L\} \bigcup G \bigcup \{D_j | l < j \le L\}$ with spatial size smaller than h_l, a convolution operation with a target feature depth of d_l is first applied before a bilinear up-sampling with a target spatial size of h_l; (iii) we fuse the processed input features of the same depth and size as a fused feature T_l by a weighted sum:

$$T_l = \sum_{m \in M_l^+} w_{m,l} bn(conv(f_{\downarrow}(m))) + \sum_{m \in M_l^-} w_{m,l} bn(f_{\uparrow}(conv(m))) \quad (1)$$

where $f_{\downarrow}()$ and $f_{\uparrow}()$ denote the bilinear down-sampling and bilinear up-sampling, respectively; $conv()$ is the 2D convolution with 3×3 kernel size; $w_{m,l}$ is the weight of the connection from input feature m to decoder stage feature D_l; and $bn()$ represents the batch normalization. $w_{m,l} = 0$ indicates the connection does not exist, while $w_{m,l} = 1$ indicates the connection exists. $bn()$ ensures the output value of $conv(f_{\downarrow}())$ or $f_{\uparrow}(conv())$ not affecting the connection importance represented by the summation weight $w_{m,l}$. Finally, decoder stage feature

D_l is obtained as $D_l = ReLU6(bn(conv(ReLU6(T_l))))$ where $ReLu6()$ is the nonlinear activation function.

To output a phase map with continuous phase values and the same image size as the input interferogram in a single channel, a regression head is added to D_1, i.e., the end feature of the decoder. The regression head is comprised of three consecutive functions: (i) 2D convolution to compress the feature depth of D_1 from 8 to 1; (ii) interpolation to make the feature spatial size of D_1 identical to the input interferogram (originally h_1 is only half the size of the input interferogram); (iii) *ReLu6* nonlinear activation and dividing the obtained value by 6 to make the output have pixel value ranging from 0 to 1 for training.

B. Deriving NAS-PRNet From Super-PRNet

The search for the optimal encoder and decoder connections for NAS-PRNet is formulated as a problem of finding the optimal binary subset of the weight set $W = \{w_{m,l} | m \in M_l^+ \bigcup M_l^-, 1 \le l \le L\}$. Considering the tradeoff between efficiency and accuracy, the optimization objectives include: (i) making the connectivity as sparse as possible to reduce the computation latency of this network; and (ii) decreasing the phase retrieval loss as much as possible. As it is computationally inefficient to search W in a discrete search space, we relax all weights $w \in W$ to be continuous, ranging from 0 to 1, to allow for gradient descent to optimize the connection weights and conduct the architecture search. The synthesized loss function $Loss_*$ used in the training process is formulated as:

$$Loss_* = Loss_t + \alpha Loss_b + \beta Loss_s, \tag{2}$$

where $Loss_t$ is the phase reconstruction loss which is calculated as the Mixed Gradient Error (MixGE) between the ground truth and the network output as defined in [20]; $Loss_b$ and $Loss_s$ are the binary loss and the network sparsity loss, respectively; and α and β are the coefficients for $Loss_b$ and $Loss_s$. $Loss_b$ and $Loss_s$ are defined as:

$$Loss_b = \frac{\sum_{w^* \in W}(-w^* log(w^*) - (1-w^*) log(1-w^*))}{len(W)}, \tag{3a}$$

$$Loss_s = \sum_{w^* \in T} w^* \tag{3b}$$

In Eq. 3a, the loss term $-wlog(w) - (1-w)log(1-w)$ in binary loss will push the weight w close to 0 or 1 during the network training process [16], and $len()$ takes the length of W. The mean of all weights $w \in W$ serves as the sparse loss as described in Eq. 3b. A smaller $Loss_s$ indicates that more weight w values are closer to 0, namely the connectivity in super-PRNet will be sparser. After tuning, we set $\alpha = 5 \times 10^{-3}$ and $\beta = 5 \times 10^{-5}$.

We first trained the super-PRNet on an NIH/3T3 cell dataset (for dataset and training details, refer to Supplementary Material). The training process took 4 hours on a Supermicro GPU server (Intel Xeon Silver 4210R CPU [×2], Nvidia RTX A6000 48GB [×1]). Then, to prune super-PRNet to get NAS-PRNet (Fig. 3), we selected the connection weight set W in the checkpoint with the best validation PSNR and SSIM. The pruning rules are as follows: (i) drop all the connections with weights $w < 0.001$; and (ii) drop all decoder stages without any input features, as well as decoder stages whose features are not used by any decoder stages. After pruning, the number of connections in super-PRNet (Fig. 3(a)) has been significantly reduced from 100 to 42, i.e., reducing the connections by 58%, to achieve NAS-PRNet that has a sparse connectivity (Fig. 3(b)). In NAS-PRNet, as both low-level and high-level features take part in the formation of each decoder, an optimal fusion of these features can ensure accurate phase retrieval.

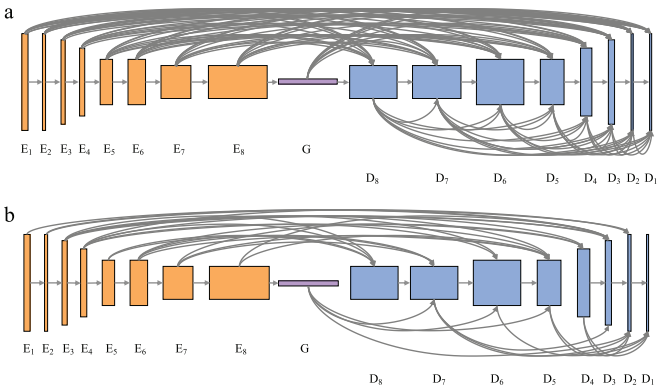

Fig. 3. (a) Connections in super-PRNet. (b) Connections in NAS-PRNet. Orange, blue, and purple rectangles represent encoder stages (E), decoder stages (D), and the ground stage (G), respectively.

In addition to applying the found connection scheme, NAS-PRNet is modified from super-PRNet in producing the fused features by removing w and $bn()$ in Eq. 1. Therefore, the fused features T_l' of NAS-PRNet are:

$$T_l = \sum_{m \in M_l^+} w_{m,l} conv(f_{\downarrow}(m)) + \sum_{m \in M_l^-} w_{m,l} f_{\uparrow}(conv(m)) \tag{4}$$

C. Evaluation of NAS-PRNet

To evaluate the phase retrieval accuracy, we trained NAS-PRNet following the same protocol as super-PRNet but only used the phase reconstruction loss $Loss_t$. We tested the phase retrieval time on a mid-range Lenovo laptop (Intel i7-9750H CPU [×1], Nvidia GeForce RTX 2060 6GB [×1]) using full NIH/3T3 cell images with size of 1024×1024 pixels.

III. Results

We compare the performance of super-PRNet and NAS-PRNet against several baseline methods, including the classic U-Net [17], the recent Mamba-UNet and EMCAD based on popular Mamba and attention mechanisms from natural language processing, respectively [18], [19], as well as SparseMask (implementation details can be found in Supplementary Material). We present the testing results using the NIH/3T3 cell dataset on a mid-range laptop in Tab. I and Fig. 4. To analyze the results, we divide the baseline methods into heavy-weight (U-Net and Mamba-UNet) and light-weight (EMCAD and SparseMask) groups. The heavy-weight group achieves higher accuracies (36.8 – 37.8 dB PSNR) but significantly slower inference speed (373 – 481 ms), while the light-weight group offers faster inference speed (21 – 88 ms) with lower accuracy (30.1 – 31.8 dB PSNR). Note that the longer inference time of Mamba-UNet with lower FLOPs is likely due to less optimized GPU implementations compared to U-Net's well-established convolution operators. In contrast, NAS-PRNet achieves an optimal balance between accuracy (36.7 dB PSNR) and efficiency (31 ms latency), thus demonstrating comparable performance with 12× speedup over

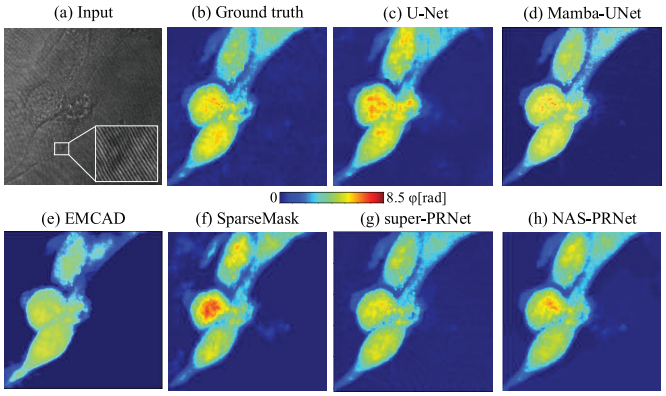

Fig. 4. Comparison of super-PRNet and NAS-PRNet with baseline methods on the NIH/3T3 cell dataset. (a) Input interferogram. (b) Ground truth. (c) U-Net. (d) Mamba-UNet. (e) EMCAD. (f) SparseMask. (g) super-PRNet. (h) NAS-PRNet.

TABLE I
PHASE RETRIEVAL PERFORMANCE COMPARISON

Methods	PSNR	SSIM	Params	FLOPs	Latency
U-Net	36.8 dB	87.4%	37.7M	971.4G	373 ms
Mamba-UNet	37.8 dB	87.8%	19.1M	94.6G	481 ms
EMCAD	30.1 dB	60.4%	3.9M	30.2G	88 ms
SparseMask	31.8 dB	65.6%	1.8M	6.9G	21 ms
super-PRNet (ours)	37.6 dB	85.6%	7.2M	17.5G	60 ms
NAS-PRNet (ours)	36.7 dB	86.6%	4.4M	11.3G	31 ms

U-Net and slightly lower performance with 15× speedup over Mamba-UNet. Moreover, the comparable accuracy between NAS-PRNet and super-PRNet demonstrates that our searching strategy is effective, as it has selectively pruned redundant connections and preserved critical connections to maintain a high phase retrieval accuracy. To test the robustness of NAS-PRNet, we used a white blood cell dataset [21] acquired from a different off-axis QPI system and achieved a high phase retrieval accuracy (PSNR 44.0 dB and SSIM 93.4%) that is comparable to U-Net (refer to Supplementary Material for details).

Note that when using the traditional Fourier transform-based phase retrieval method (i.e., the ground truth map), phase unwrapping is required after obtaining the calibrated phase map. Using the Goldstein algorithm, phase unwrapping takes 528 ms for an image size of 1024×1024 when executed on the CPU of the laptop. In contrast, our NAS-PRNet outputs the unwrapped phase in just 31 ms on the same laptop using GPU, which is over 17 times faster. The phase unwrapping range of NAS-PRNet is currently limited to 0 - 12 rad (the phase range in the dataset). However, this range can be extended by training the network with datasets containing larger phase values. Moreover, NAS-PRNet can automatically correct system aberration without using a calibration phase map, which significantly simplifies the phase imaging experiments.

IV. CONCLUSION

In conclusion, we have developed NAS-PRNet for phase retrieval and optimized its architecture to balance the output accuracy and inference speed. Compared with the recent Mamba-UNet, NAS-PRNet reduces inference time by 15× with only slightly lower phase retrieval accuracy. With the high phase retrieval speed offered by NAS-PRNet and the single-shot capability of off-axis QPI, one may demonstrate many real-time imaging applications, such as profiling the morphology of living cells and quantifying their dynamics. The current search space of NAS-PRNet is only limited to the connection scheme, but it could be further expanded to cover layer depth, layer manipulation, and so on, which may lead to the discovery of a more efficient network architecture.

REFERENCES

[1] Y. Park, C. Depeursinge, and G. Popescu, "Quantitative phase imaging in biomedicine," *Nature Photon.*, vol. 12, no. 10, pp. 578–589, 2018.
[2] M. Niu, Y. Wang, L. Duan, R. Sun, and R. Zhou, "Compact morpho-molecular microscopy for live-cell imaging and material characterization," *Laser Photon. Rev.*, vol. 18, no. 4, Apr. 2024, Art. no. 2300741.
[3] T. Ling et al., "High-speed interferometric imaging reveals dynamics of neuronal deformation during the action potential," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 19, pp. 10278–10285, May 2020.
[4] M. Takeda, H. Ina, and S. Kobayashi, "Fourier-transform method of fringe-pattern analysis for computer-based topography and interferometry," *J. Opt. Soc. Amer.*, vol. 72, no. 1, pp. 156–160, 1982.
[5] P. Ferraro et al., "Compensation of the inherent wave front curvature in digital holographic coherent microscopy for quantitative phase-contrast imaging," *Appl. Opt.*, vol. 42, no. 11, pp. 1938–1946, 2003.
[6] R. M. Goldstein, H. A. Zebker, and C. L. Werner, "Satellite radar interferometry: Two-dimensional phase unwrapping," *Radio Sci.*, vol. 23, no. 4, pp. 713–720, Jul. 1988.
[7] H.-Y. Chen, S.-H. Hsu, W.-J. Hwang, and C.-J. Cheng, "An efficient FPGA-based parallel phase unwrapping hardware architecture," *IEEE Trans. Comput. Imag.*, vol. 3, no. 4, pp. 996–1007, Dec. 2017.
[8] H. Pham, H. Ding, N. Sobh, M. Do, S. Patel, and G. Popescu, "Off-axis quantitative phase imaging processing using CUDA: Toward real-time applications," *Biomed. Opt. Exp.*, vol. 2, no. 7, pp. 1781–1793, 2011.
[9] K. Wang et al., "On the use of deep learning for phase recovery," *Light, Sci. Appl.*, vol. 13, no. 1, p. 4, Jan. 2024.
[10] T. Chang, D. Ryu, Y. Jo, G. Choi, H.-S. Min, and Y. Park, "Calibration-free quantitative phase imaging using data-driven aberration modeling," *Opt. Exp.*, vol. 28, no. 23, pp. 34835–34847, 2020.
[11] Y. Yao, X. Shu, and R. Zhou, "Deep learning based phase retrieval in quantitative phase microscopy," *Proc. SPIE*, vol. 11351, Apr. 2020, Art. no. 113510W.
[12] K. Wang, J. Dou, Q. Kemao, J. Di, and J. Zhao, "Y-Net: A one-to-two deep learning framework for digital holographic reconstruction," *Opt. Lett.*, vol. 44, pp. 4765–4768, Oct. 2019.
[13] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *J. Mach. Learn. Res.*, vol. 20, no. 1, pp. 1997–2017, 2019.
[14] P. Ren et al."A comprehensive survey of neural architecture search: Challenges and solutions," *ACM Comput. Survey*, vol. 54, no. 4, pp. 1–34, 2021.
[15] C. Liu et al., "Auto-DeepLab: Hierarchical neural architecture search for semantic image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 82–92.
[16] H. Wu, J. Zhang, and K. Huang, "SparseMask: Differentiable connectivity learning for dense image prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6767–6776.
[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, 2015, pp. 234–241.
[18] Z. Wang, J.-Q. Zheng, Y. Zhang, G. Cui, and L. Li, "Mamba-UNet: UNet-like pure visual mamba for medical image segmentation," 2024, *arXiv:2402.05079*.
[19] M. M. Rahman, M. Munir, and R. Marculescu, "EMCAD: Efficient multi-scale convolutional attention decoding for medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 11769–11779.
[20] S. Van Der Jeught, P. G. G. Muyshondt, and I. Lobato, "Optimized loss function in deep learning profilometry for improved prediction performance," *J. Phys., Photon.*, vol. 3, no. 2, Apr. 2021, Art. no. 024014.
[21] X. Shu et al., "Artificial-intelligence-enabled reagent-free imaging hematology analyzer," *Adv. Intell. Syst.*, vol. 3, no. 8, Jun. 2021, Art. no. 2000277.