IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

A Voxelized Transformer-based Neural Network for 3D Reconstruction from Multi-Energy SEM Backscattered Electrons

CAIZHI ZHENG[1], RONGHAN HONG[2,3], HAO-JIE HU[1] and Qing Huo Liu[3,1], (Fellow, IEEE)

[1]Institute of Electromagnetics and Acoustics, Fujian Provincial Key Laboratory of Electromagnetic Wave Science and Detection Technology, Xiamen University, Xiamen 361005, China.
[2]Tangshan Technology Company, Ningbo 315000, China.
[3]School of Electronic Science and Technology, Eastern Institute of Technology, Ningbo 315200, China.

Corresponding author: Qing Huo Liu (qhliu@eitech.edu.cn)

ABSTRACT Three-dimensional (3D) data analysis, especially 3D reconstruction, is critical within integrated circuits for quality control and inspection. The traditional 3D reconstruction, like multi-energy deconvolution scanning electron microscopy (MED-SEM), can achieved resolution of 5 nm in the horizontal direction and 10 nm in the vertical direction from backscattered electron images of objects at different depth layers. However, the traditional methods are difficult to obtain high-precision reconstruction results in the longitudinal direction, and the reconstruction results of most methods contain a large number of artifacts. Thus, in this work, a hybrid CNN-Transformer network, called BSE-VoxNets, is proposed for 3D high precision reconstruction. The proposed BSE-VoxNets consists three parts: transformer feature extractor, convolutional upsampling block and feature fusion block. 10000 BSE samples were used for training BSE-VoxNets, with 200 samples for testing, and the optimal cross-union ratio (IoU) of the test dataset reached 0.953. Two numerical cases are used to respectively verify the inversion performance of the proposed method for different depths and different feature scales. The reconstruction results show that the proposed BSE-VoxNets can reach a resolution of 2 nm in both the horizontal and vertical dimensions.

INDEX TERMS 3D reconstruction, deep learning, machine learning, scanning electron microscopy.

I. INTRODUCTION

THREE-DIMENSIONAL (3D) data analysis, particularly 3D reconstruction, has become a critical research area in integrated circuit inspection and reverse engineering. A landmark advancement in this field is ptychographic X-ray computed tomography (PXCT) [1], which enables non-destructive high-resolution 3D imaging. However, PXCT requires access to X-ray synchrotrons capable of producing the highly coherent, intense X-ray beams essential for scanning probe microscopy [2]. These stringent source requirements significantly limit its practical adoption.

To overcome these limitations, this study explores the potential of achieving comparable 3D reconstruction quality using established platforms such as scanning electron microscopy (SEM). Unlike synchrotron X-rays, SEM systems operate at much lower electron energies, limiting penetration depth [3], [4]. Consequently, destructive techniques like focused ion beam (FIB) tomography are often required for accurate 3D reconstructions. Reconstructing 3D models from tomographic 2D projections further demands advanced algorithms for offset correction and alignment [5]. For instance, Sorzano *et al.* [6] stabilized angular alignment globally using Wiener filtering and Fourier-space reconstruction, while Bogensperger *et al.* [7] developed a marker-free technique that jointly aligns and reconstructs images without fiducial markers.

Recent advances in computer vision and deep learning have revolutionized 3D reconstruction methodologies [8]–[10]. In tomographic imaging, deep learning models increasingly replace traditional alignment strategies, streamlining workflows and improving accuracy. For example, in scanning transmission electron microscopy (STEM), a Neural Radiance Fields (NeRF)-based model [11] simultaneously learns the 2D sensor noise distribution and reconstructs an implicit 3D volumetric representation, eliminating the need for separate noise modeling and reconstruction stages. Another approach by F.P. de Isidro-Gómez *et al.* [12] leverages convolutional neural networks (CNNs) to analyze fiducial marker characteristics in projection images, enabling automated detection and correction of misalignment artifacts. In

semiconductor metrology, domain-specific challenges such as repetitive nanostructures and low signal-to-noise ratios in SEM images have driven innovations in unsupervised learning. For instance, Houben *et al.* [13] proposed an unsupervised domain adaptation framework that aligns SEM image features with physical surface topography, enabling precise reconstruction of nanoscale fin structures without requiring labeled training data. These advancements highlight the versatility of deep learning in addressing both general and domain-specific bottlenecks in 3D reconstruction, from noise-robust reconstruction to artifact-free alignment in resource-constrained imaging environments.

In contrast to destructive sample preparation techniques, recent work by [14] proposes a nondestructive alternative: modulating the electron beam energy to capture multi-depth images, followed by deconvolution algorithms to improve resolution and fidelity. Michiel *et al.* [15] advanced this approach with the Multi-Energy Deconvolution SEM (MED-SEM) algorithm, which processes multi-energy backscattered electron (BSE) images to achieve high-quality 3D reconstructions.

Despite its promise, MED-SEM faces two key challenges. First, the fidelity of 3D reconstruction critically depends on BSE point spread functions (PSFs), which govern the resolution and clarity of raw SEM images. While MED-SEM uses blind source separation to devolve signals across energy levels—indirectly estimating PSFs—selecting accurate PSFs for diverse energy slices and 3D structures remains highly complex. Second, multi-energy images generated via single-tilt acquisition are prone to artifacts that degrade reconstruction quality.

Compared with the previous work, the contribution of this work as follow: (1) A hybrid CNN-Transformer network, called BSE-VoxNets, is proposed for sub-5 nm resolution 3D reconstruction directly from multi-energy SEM data, which is the first deep learning solution enabling artifact-free to our knowledge. The proposed BSE-VoxNets can achieve an average loU of 0.953 on 200 test geometries, with stable performance across varying depths and scales. (2) Compared with the reconstruction resolution of MED-SEM, the proposed BSE-VoxNets can achieve 2 nm resolution in both lateral and vertical dimensions for 3D reconstruction. (3) By integrating transformer-based global attention and depth-aware convolutional upsampling, the proposed method can resolve PSF ambiguities and eliminate depth-dependent artifacts.

The subsequent sections of this article are organized as follows. Section II provides a comprehensive description of the method, meticulously delineated into two components: dataset composition and network architectural configurations. Section III presents the results of BSE-VoxNets across two distinct core-shell scenarios, both compare analyses utilizing the MED-SEM approach. To the best of our knowledge, this work is new in employing deep learning based method in multi-energy SEM reconstruction. The study provides an effective multi-energy SEM reconstruction method. This approach adeptly addresses two key challenges of PSFs selection and artifacts elimination inherent in traditional multi-energy 3D reconstruction algorithms and significantly enhances its resolution capabilities.

II. METHODS AND DETAILS

A. DATASET

Due to the considerable time investment required to obtain BSE images through experimental methods, the acquisition of training data was pursued via simulation. There are many methods for modeling the electron scattering process via Monte Carlo stochastic simulation [16]–[18]. The Nebula simulator [19] is employed in this work for simulation purposes, utilizing first-principle models and accelerated significantly by GPUs. The Nebula simulator is capable of capturing electron emissions that originate from a variety of factors, including the geometric configuration, the initial electron attributes, and the material properties, as shown in Fig. 1. The geometric configuration is depicted as a semiconductor's surface, articulated through a series of triangular facets. The initial electron attributes denote the baseline electron characteristics at the onset of the simulation process, such as initial energy. The material properties encompass data pertaining to scattering cross-sectional characteristics of the material.

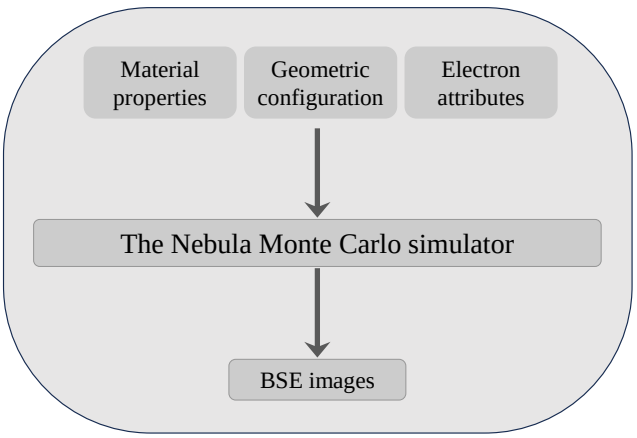

FIGURE 1. The specific flow of the Nebula simulator.

A comprehensive dataset comprising 10,000 geometric configurations was generated using MATLAB.The dimensions of each configuration were specified as 20 units in the X-direction, 20 units in the Y-direction, and 26 units in the Z-direction, with 1 nm allocated for each unit. From top to bottom, the backscattered electrons detector was set at $z = 0$ nm, and electron beam emission was set at $z = 2$ nm. The exterior silicon (Si) was set at a depth of $z \in [5, 25]$ nm, with the internal copper (Cu) material situated within, and reflective boundary conditions established around the edges. The terminator, which eliminates all electrons with inward movement was set at $z = 26$ nm. The remaining volume was filled with vacuum. Subsequently, the Nebula simulator was employed to produce BSE images across 20 discrete

energy levels, uniformly incremented by 0.5 KeV, ranging from 0.5 KeV to 10 KeV. This resulted in the generation of the multi-energy BSE (ME BSE), which was employed as the training and testing dataset.The dataset with mapping relationships is obtained by identifying patterns to be used as the output and BSE images to be used as the input. This methodology establishes an end-to-end structured dataset that correlates the initial geometric configurations with the resulting BSE images, thereby providing a foundation for training a deep learning network model to accurately predict material-specific electron backscatter patterns. The specific flow of the Nebula simulator and the specific simulation geometry with the simulation setup (XOY surface profiles) of a Cu-Si core-shell nanowire example are presented in Fig. 2. A representatively random sample of 2% of the data was designed as the test dataset, which is excluded from the training process.

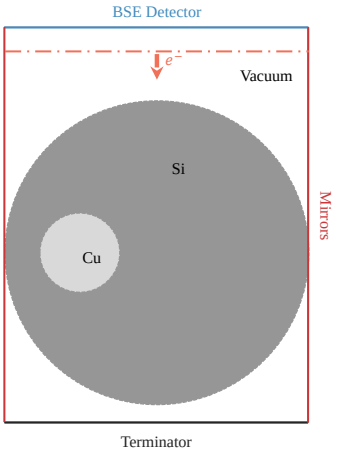

FIGURE 2. The specific simulation geometry with the simulation setup of a Cu-Si core-shell nanowire example in XOY surface profiles.

B. DEEP LEARNING ARCHITECTURE

For the task of inner 3D structure reconstruction, the BSE-VoxNets network model is proposed. The design of BSE-VoxNets is guided by the need to efficiently capture both local and global features from limited datasets, while effectively bridging the gap between 2D input images and the 3D volume prediction demands. The network architecture is composed of three principal and logically connected modules: a 2D image feature extractor, a reduced upsampling module, and a 2D-to-3D feature extractor. This configuration is illustrated in Fig.3.

The initial module is based on the S2MLPv2 [20], which employs split attention to aggregate the three distinct spatial shift feature map groups to obtain channel features for hierarchical reconstruction. This strategy leverages the strengths of MLP-based models in modeling non-local dependencies, while enhancing feature integration across various spatial perspectives through attention mechanisms. This is particularly crucial for extracting minute features that lie beyond the reach of CNN's local receptive fields. In order to enhance the performance of transformer backbone models on small datasets, the sequential overlapping patch embedding (SOPE) proposed by Lu *et al.* [21] have been implemented. In comparison to other prevalent patch embedding modules [8], [9], SOPE incorporates two affine transformation layers at the beginning and end, which may serve to obviate discontinuities resulting from chunking and ensure the retention of salient low-level features. Following the initial processing of the ME BSE through the SOPE and N1 S2MLPv2 blocks, a new iteration of the SOPE and N2 S2MLPv2 blocks has been employed, as the use of smaller patches has been observed to yield superior outcomes.

The response of the second module is to restore the embedding patch back to the image scale. To accelerate the training process, the standard upsampling layers with a convolutional feed forward submodule (ConvFFN) is utilized. In the ConvFFN, a depthwise convolutional layer and an identity skip connection are employed to learn features on each channel independently while preserving gradients in the deeper regions of the model. The 2D convolutional layer is substituted for the linear transformation layer utilized in the feed forward operation by the original Transformer [22], as the global linear transformation would otherwise disrupt the image scale features after the feature extractor has performed an upsampling. Three instances of a combination of upsampling and ConvFFN are utilized.

The final module is a simple 2D convolutional layer with a 3D convolutional layer, resulting in a final 3D geometric output. To obtain discrete binary classification results directly in , the one-hot encoding method [23] for label and the net output layer is employed.

Weighted binary cross-entropy (WBCE) is used as the loss function during model training:

$$L = -\sum_{i=1}^{N}[\lambda \cdot y_i \log h(y_i) + (1-\lambda)\cdot(1-y_i)\log\left(1-h(y_i)\right)] / N \tag{1}$$

where y_i represents the ground truth (GT) in labels, which has two discrete values: $y_i = 0$ *or* 1 indicates Si (negative) or Cu (positive); the hyperparameter λ serves to adjust the contribution of negative labels and positive labels; and $h(y_i)$ represents the probability of the model predicting (Pred) y_i. The WBCE loss between the predicted probability distribution and one-hot encoded ground truth at each voxel retains all probability information required for effective model optimization. After initial training with the WBCE loss function, the Lovász loss function, proposed by Berman [24], based on the intersection over union (IoU) estimation is employed:

$$IoU = \frac{GT \cap Pred}{GT \cup Pred} \tag{2}$$

to fine-tune the model.

Unlike pure CNN or purely Vision Transformer designs, BSE-VoxNets utilizes a two-stage SOPE and spatial shift MLP blocks, which are better suited for handling diverse feature scales and spatial relationships within 2D image stacks before producing a 3D output. The BSE-VoxNets architecture integrates both CNNs and Transformer-based modules within

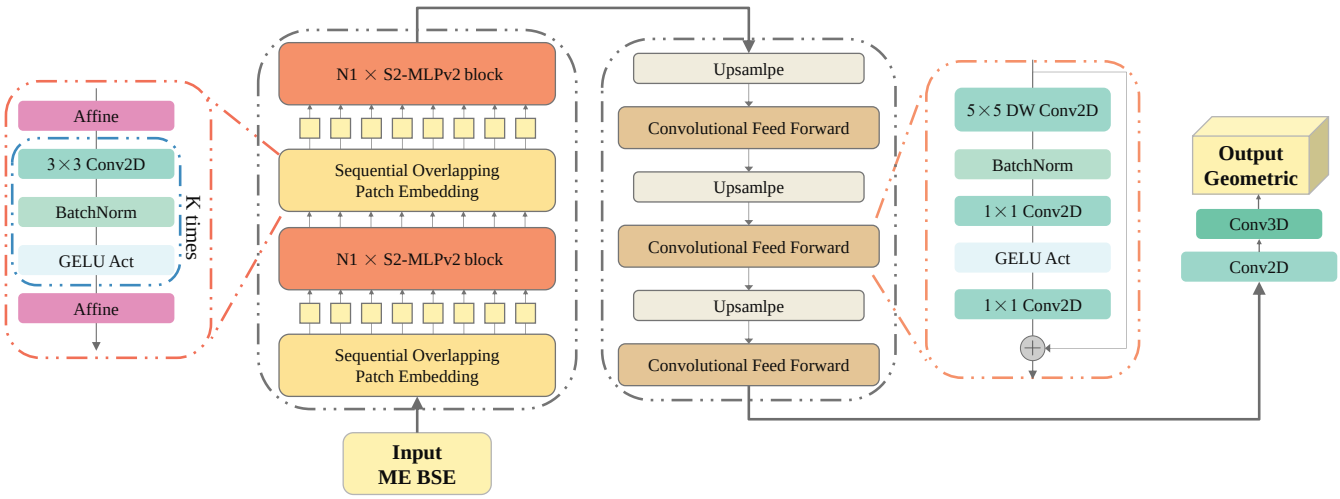

FIGURE 3. **The detailed architectural design of BSE-VoxNets model, which comprises three components from input to output: a 2D image feature extractor, a 2D reduced upsampling module, and a 2D to 3D feature extractor.**

a unified framework, specifically designed to efficiently extract spatial and channel-wise features from multi-energy SEM data.

As demonstrated in Table 1, the ablation study validates the synergistic contributions of BSE-VoxNets' core modules. To achieve a more accurate assessment, the mean relative error (MRE) is employed as the evaluation metric. The full configuration (SOPE + S2-MLPv2 + ConvFFN) achieves the lowest MRE (8.5%). Removing any component leads to performance degradation:

- Omitting SOPE increases MRE to 9.3%.
- Removing S2-MLPv2 dramatically degrades performance with MRE rising to 38.6%.
- Substituting ConvFFN with a standard FFN degrades MRE to 14.8%.

These results quantitatively confirm that the hybrid architecture's global-local interaction mechanisms are essential for achieving nanoscale reconstruction fidelity.

TABLE 1. **Ablation study results of each module in BSE-VoxNets network**

SOPE	S2-MLPv2	ConvFFN	FFN	MRE
√	√	√	×	8.5%
×	√	√	×	9.3%
√	×	√	×	38.6%
√	√	×	√	14.8%

Multiple configurations of BSE-VoxNets have been evaluated, designated as Lite, Small, Medium, and Moderate, encompassing a range of network block sizes, as detailed in Table 2. In the process of model inference, the detailed test dataset model predicting results is illustrated in Fig. 4. The model is tested once in every five epochs, with a total of 250 epochs. The Lite setting model is terminated at the 205th epoch due to the application of an early stop training strategy, which indicates that the loss of the model did not decrease over 100 epochs. Given that the Moderate setting is the most optimal, it is therefore adopted as the practical use model.

TABLE 2. **Detailed information and test results of the four main configurations**

Settings	Lite	Small	Medium	Moderate
1st SOPE patch size	4×4	4×4	4×4	4×4
1st Hidden size	192	192	384	384
N1 of Blocks	2	4	7	10
2nd SOPE patch size	2×2	2×2	2×2	2×2
2nd Hidden size	384	384	768	768
N2 of Blocks	5	14	17	17
Total parameters	16M	36M	173M	179M
Floating point operations (FLOPs)	1.1B	1.2B	5B	5.1B
Mean Intersection over Union (mIoU)	0.796	0.826	0.935	0.953

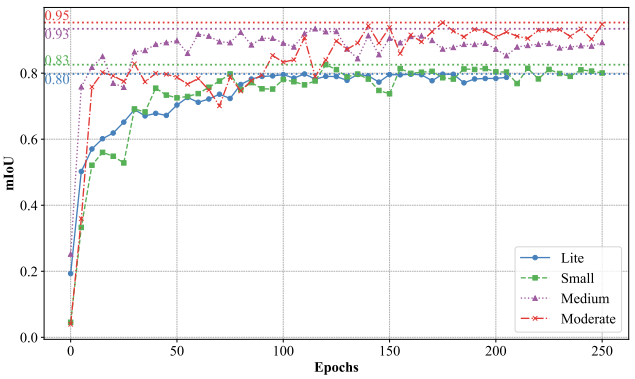

FIGURE 4. **mIoU comparision of four configurations of BSE-VoxNets in Test dataset.**

III. RESULTS

The results of the BSE-VoxNets deep learning approach are presented in two sections. The first part encompasses the

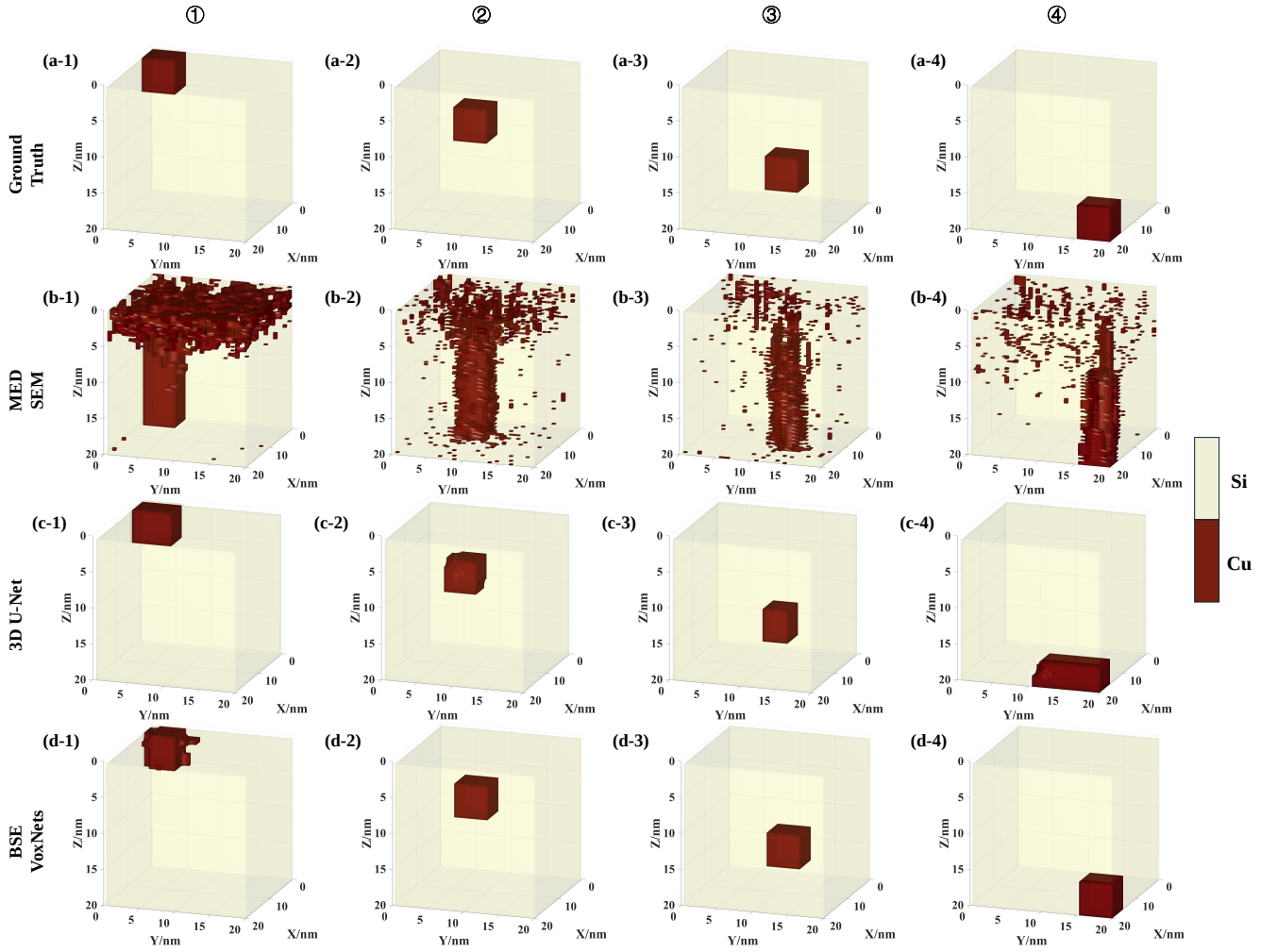

FIGURE 5. **The 3D reconstruction results for core-shell nanocube numerical example of MED-SEM, 3D U-Net and BSE-VoxNets: (a) the ground truth of core-shell nanocube with core cube in four positions, (b) the reconstruction results of MED-SEM, (c) the reconstruction results of 3D U-Net, (d) the reconstruction results of BSE-VoxNets.**

evaluation of the accuracy and the pivotal depth information of BSE-VoxNets's reconstruction of the fundamental geometry, utilizing core-shell nanocube at varying depths. The subsequent part involves the assessment of the limiting resolution and the reliability of the reconstruction of analogous structures by BSE-VoxNets, employing a core-shell nanowire structure at different core radii. Both segments are compared with the results obtained from MED-SEM and 3D U-Net [25].

BSE-VoxNets is trained using the AdamW optimizer [26]. Three key hyperparameters - the initial learning rate, weight decay, and WBCE loss weighting parameter λ - are determined through grid search optimization evaluating performance on the test set and set to the test numerical examples. The grid search spans initial learning rates of $\{5\times10^{-3}, 1\times10^{-3}, 5\times10^{-4}, 1\times10^{-4}, 5\times10^{-5}, 1\times10^{-5}\}$, weight decay values of $\{1\times10^{-3}, 1\times10^{-4}, 1\times10^{-5}\}$, and λ values of $\{1/n \,|\, n = 2, 3, ..., 15\}$. Evaluating over 200 hyperparameter configurations, the optimal configuration achieving the best 0.98 mIoU is found to be an initial learning rate of 1×10^{-4},

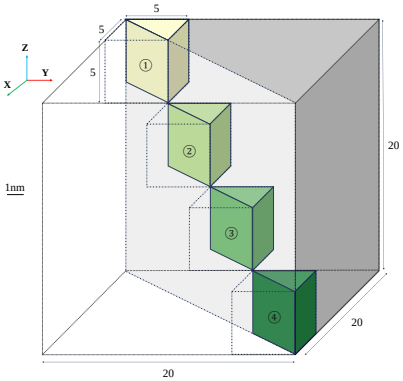

FIGURE 6. **The schematic of the core-shell nanocube numerical example with four depth values.**

weight decay of 1×10^{-4}, and λ of $1/12$ (≈ 0.0833). A cosine annealing schedule is applied to the learning rate throughout the training process. All implementations are based on Py-

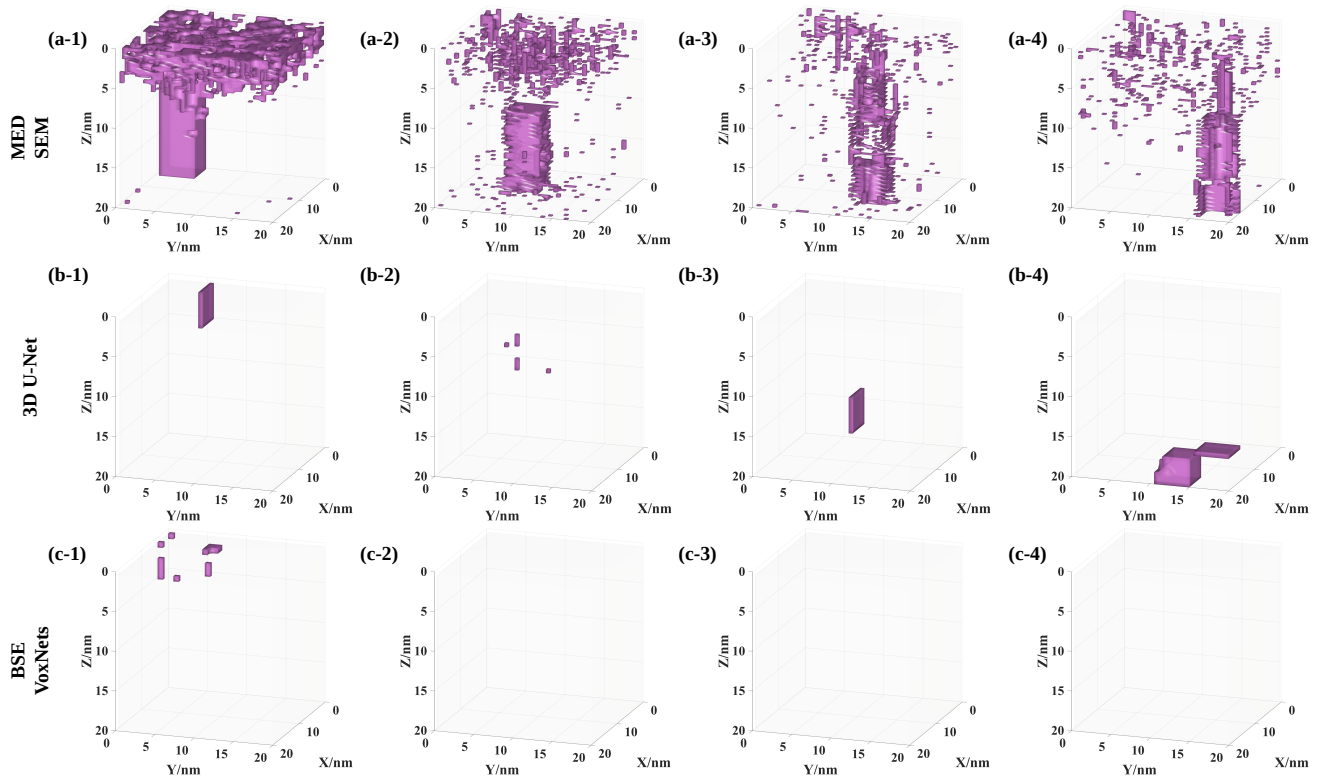

FIGURE 7. **The 3D reconstruction results error for core-shell nanocube numerical example of BSE-VoxNets and MED-SEM: (a) the absolute error between the results of MED-SEM and ground truth, (b) the absolute error between the results of 3D U-Net and ground truth, (c) the absolute error between the results of BSE-VoxNets and ground truth.**

Torch, and experiments are conducted on an NVIDIA RTX 3090 GPU.

For execution time BSE-VoxNets achieves an average inference time of 1 seconds per volume, compared to 300 seconds for MED-SEM (default setting). These results indicate that BSE-VoxNets offers competitive efficiency alongside its improved reconstruction accuracy.

A. CORE-SHELL NANOCUBE IN VARYING DEPTHS RECONSTRUCTION

In this numerical example, four cases were proposed to evaluate the performance of BSE-VoxNets, and each cases includes four distinct core-shell nanocube structures. Each structure consists of a 20-nanometer Si shell encapsulating a 5-nanometer Cu core, with the cores positioned sequentially along the diagonal axis. While all four core-shell structures share identical geometric complexity, the sequential diagonal positioning introduces progressive attenuation of spatial signal distinctiveness at deeper layers. This arrangement allows assessment of the model's ability to reconstruct basic geometrical configurations and accurately distinguish between different materials at varying depths.

The schematic representation of these nanocubes is provided in Fig. 6, which visually outlines the spatial distribution of the core and shell materials. The corresponding 3D reconstruction results and the results error from ground truth are displayed in Fig. 5 and Fig. 7, showcasing the model's output in comparison to the actual structures. The IoU measurements for each reconstructed sample are tabulated in Table 3. Given the nature of the MED-SEM method, which outputs probability values ranging from 0 to 1, thresholding techniques were applied to optimize the segmentation of reconstructed images. Threshold values between 0.5 and 0.9 were applied to discern the material boundaries. Analysis showed that a threshold of 0.5 was most effective for the first position, 0.7 for the second, and 0.9 for the subsequent positions.

TABLE 3. **The comparative IoU results for MED-SEM, 3D U-Net and BSE-VoxNets for the core-shell nanocube numerical example of four depth values.**

Position	①	②	③	④
IoU of MED-SEM	0.0595	0.165	0.206	0.204
IoU of 3D U-Net	0.833	0.952	0.800	0.465
IoU of BSE-VoxNets	0.902	1	1	1

As is depicted in Fig. 5, the reconstruction results of MED-SEM have surface topography artifacts referred in the original article cannot be ignored, but deep learning methods achieve considerable success. For 3D U-Net, although it performed well, with IoU greater than 0.8 in most cases, it showed a significant decrease at "Position ④" (0.465). This performance discrepancy may be attributed to the inherent limitations of fully convolutional networks in handling depth-dependent

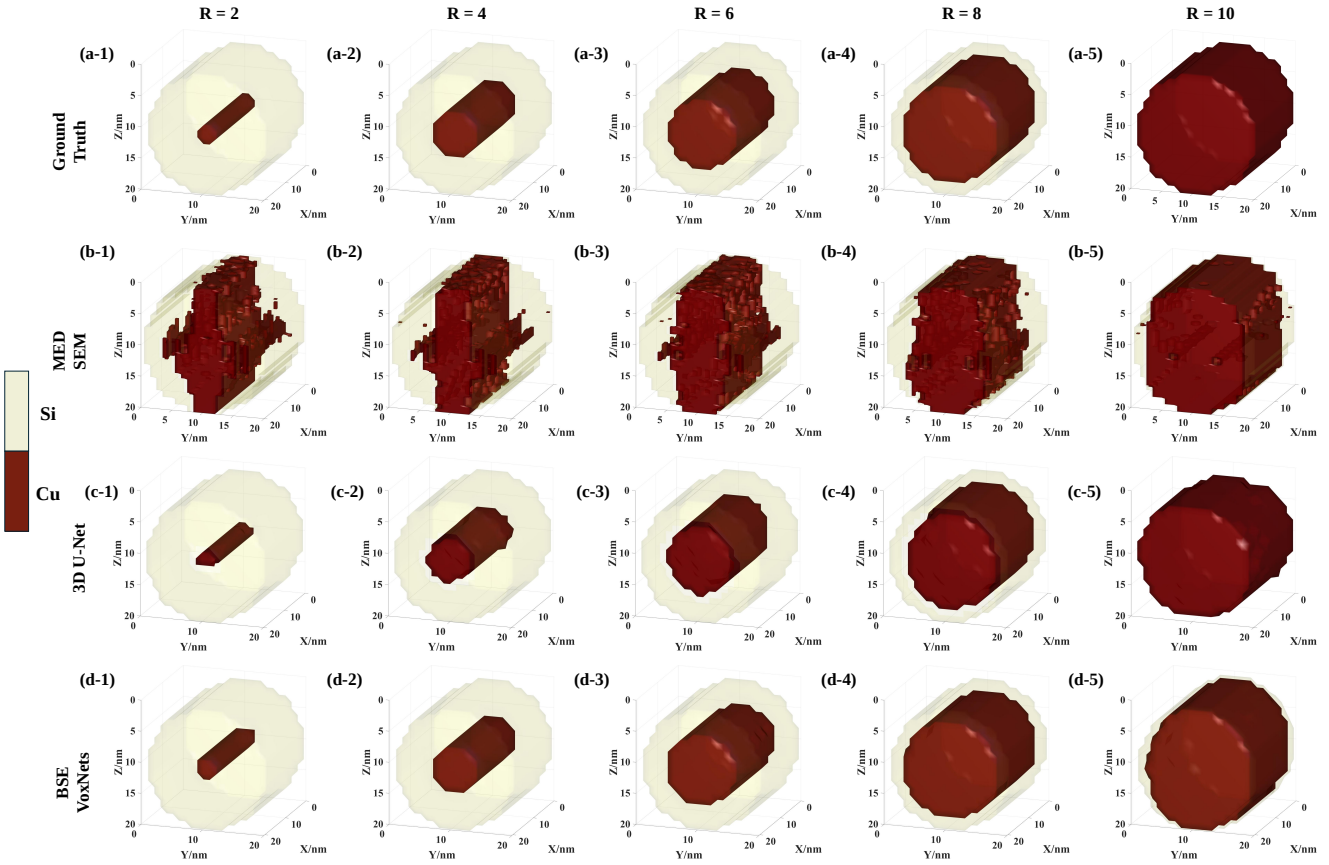

FIGURE 8. **The 3D reconstruction results for core-shell nanowire numerical example of MED-SEM, 3D U-Net and BSE-VoxNets: (a) the ground truth of core-shell nanocube with core cube in four positions, (b) the reconstruction results of MED-SEM, (c) the reconstruction results of 3D U-Net, (d) the reconstruction results of BSE-VoxNets.**

feature degradation.

The IoU trajectory of 3D U-Net reveals limitations in handling complex geometries. The local receptive fields of convolutional layers struggle to maintain material contrast across compressed latent spaces. This leads to loss of critical boundary information needed for ultra-thin core-shell interface reconstruction.This demonstrates that its segmentation accuracy in some complex or special positions is not as stable as BSE-VoxNets. In contrast, BSE-VoxNets demonstrates more remarkable results with no errors observed in the last three positions. This robustness originates from its hybrid Transformer-CNN architecture, where the global attention mechanisms effectively compensate for the local receptive field limitations of convolutional networks. For the first position, BSE-VoxNets obtain the 0.902 IoU value. The observed error in this instance is attributable to the positioning of the Cu core cube at the top of the shell cube, leading to the absence of supplementary information for BSE at high energy levels (thus making this data more biased relative to the full dataset). This is also the underlying reason why MED-SEM reconstruction performed poorly for the first position.

This result demonstrates BSE-VoxNets has the ability to reconstruct the basic geometry cube and could predict the accurate depth information at varying depths.

B. CORE-SHELL NANOWIRE IN DIFFERENT RADII RECONSTRUCTION

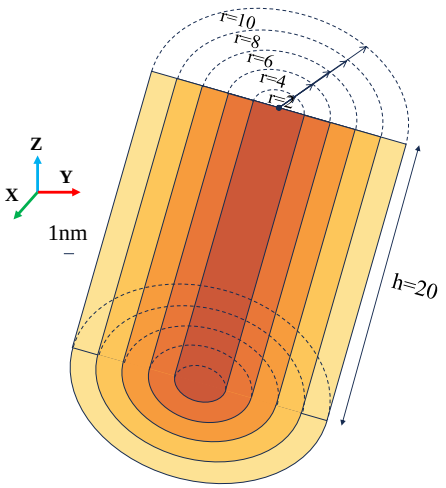

FIGURE 9. **The schematic of the core-shell nanowire numerical example with five radius values.**

This segment of analysis extends to evaluating BSE-VoxNets on a numerical example comprising five concentric

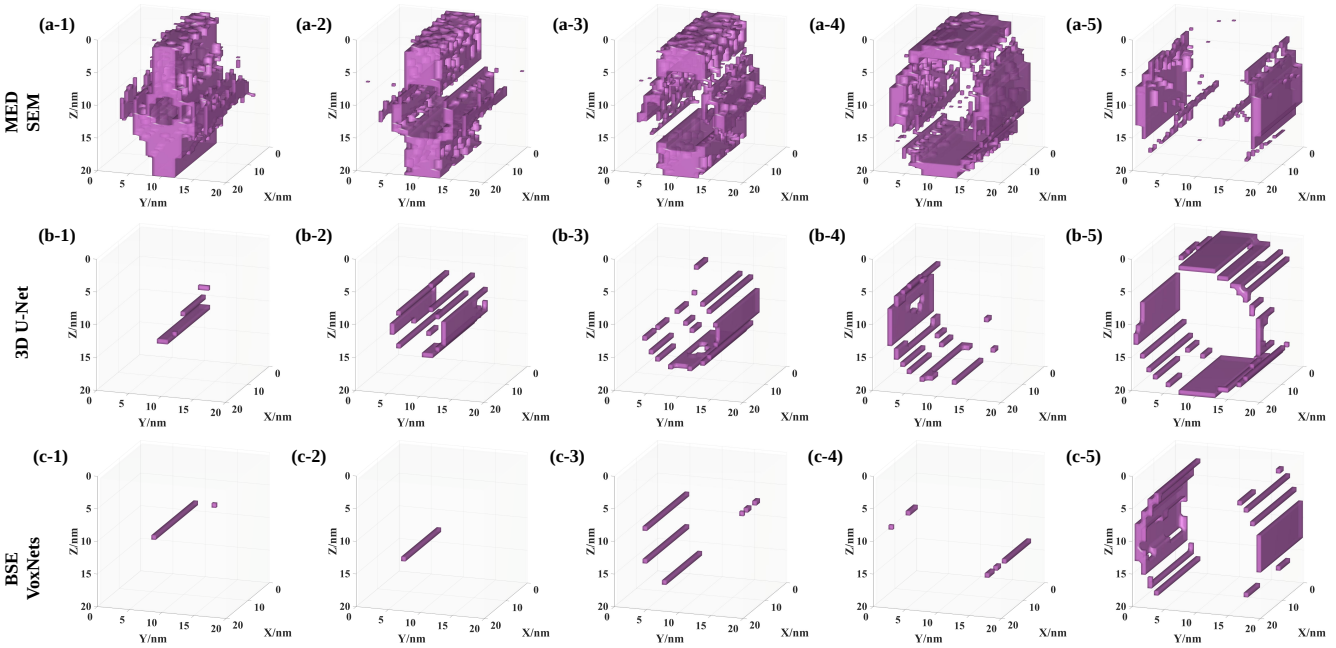

FIGURE 10. **The 3D reconstruction results error for core-shell nanowire numerical example of BSE-VoxNets and MED-SEM: (a) the absolute error between the results of MED-SEM and ground truth, (b) the absolute error between the results of 3D U-Net and ground truth, (c) the absolute error between the results of BSE-VoxNets and ground truth.**

column core-shell nanowire structures. These nanowires feature an incremental Cu radius, ranging from 2 nm to 10 nm in increments of 2 nm, encapsulated by a constant 10-nanometer Si shell. This variation in radius was employed to examine how changes in the analogous geometry affect the model's reconstruction capabilities and the resolution limit.

Fig. 9 provides a visual representation of the numerical example, while the corresponding 3D reconstruction results are displayed in Fig. 8. The IoU measurements are compiled in Table 4. For the MED-SEM method applied to this numerical example, a threshold value of 0.5 was most effective for the sample with a 10 nm Cu radius. For the remaining samples with varying radii, a threshold of 0.8 yielded the optimal results.

TABLE 4. **The comparative IoU results for for MED-SEM, 3D U-Net and BSE-VoxNets for the core-shell nanowire numerical example of five radius values.**

radius of Cu (nm)	2	4	6	8	10
IoU of MED-SEM	0.095	0.324	0.508	0.641	0.872
IoU of 3D U-Net	0.788	0.777	0.879	0.940	0.900
IoU of BSE-VoxNets	0.920	0.984	0.974	0.995	0.918

As demonstrated in the results, MED-SEM exhibits significant resolution constraints, demonstrating a pronounced inverse relationship between feature size and reconstruction fidelity. When the core radius falls below 4 nm, the reconstruction artifacts become so pronounced that both the morphology and dimensions of the Cu core become indistinguishable, with depth-related artifacts permeating the entire reconstructed volume. This aligns with its inherent 4 nm lateral/10 nm depth resolution. Meanwhile, deep learning methods have much better results. 3D U-Net achieves higher IoU at larger radii (6–10 nm), but it exhibits geometric irregularities and severe artifacts at smaller radii (2, 4 nm). This results in a significant decrease in IoU (to approximately 0.78) in these cases, demonstrating poor adaptability to variations in target size. And for BSE-VoxNets, it maintains consistently high reconstruction fidelity across all tested scales with IoU exceeds 0.9, which significantly surpasses that of MED-SEM and 3D U-Net. Compared to 3D U-Net, the integrated global attention mechanism in BSE-VoxNets effectively suppresses size-dependent artifacts while preserving nanoscale geometric regularity. Even in the lowest IoU case, which is 0.918 for the 10-nm radius, BSE-VoxNets still outperforms the best results of MED-SEM and 3D U-Net for the same sample (0.872 and 0.900, respectively). This is likely because the training dataset consists of two materials. However, when the radius of Cu is 10 nm, only a single material is present. The critical case, which fall outside the range of the training dataset, validates the generalizability of BSE-VoxNets.

This outcome demonstrates the efficacy and reliability of BSE-VoxNets's 3D reconstruction for analogous structures at disparate scales. Considering the results collectively, it is estimated that the resolution of the BSE-VoxNets reconstruction is 2 nm for both lateral and depth, which is superior to the resolution of 4 nm for lateral and 10 nm for depth with MED-SEM.

IV. DISCUSSION

BSE-VoxNets achieves a significant advancement in the field of 3D reconstruction based on multi-energy SEM BSE images, leveraging the strengths of deep learning to address two persistent challenges in traditional methods: PSFs ambiguity and artifact-induced reconstruction errors. The results show that BSE-VoxNets achieves significantly higher IoU scores in the core-shell nanocube and nanowire scenes compared to the MED-SEM algorithm and 3D U-Net, while obtaining more realistic reconstruction results.

Compared to other conventional methods, a significant advantage of BSE-VoxNets is its ability to generalize well to core-shell structures with different depths and radii. In particular, the high IoU values achieved on geometries that do not appear explicitly in the training data suggest that the model captures the spatial and material relationships in the SEM BSE data, rather than just overfitting a specific configuration. In addition, it exhibits robustness in the processing of multi-energy channel input data, reflecting the efficiency of its hybrid convolutional neural network (CNN)-Transformer backbone network in extracting and integrating multi-scale features. In addition, BSE-VoxNets significantly reduces the common artifact problem and achieves more stable and artifact-free reconstruction results through statistical relationships in large datasets.

However, there are still some points that deserve further exploration. First, BSE-VoxNets relies on simulated datasets rather than experimentally acquired SEM images. While simulated data can provide a controlled benchmarking environment and support large-scale data generation, actual SEM BSE images may contain additional noise, system-specific aberrations, or unmodeled physical effects. The ability of the current model to generalize real experimental data needs to be validated in future studies. Second, the current training and testing involves binary material classification (Si and Cu), more complex multi-component or gradient composition material systems may pose new challenges. The ability to extend the model to handle more than two materials, irregular or layered structures, and variable boundary conditions would significantly increase its utility. In future research, the above two issues will be improved and refined.

V. CONCLUSIONS

In this paper, a deep learning-based 3D reconstruction algorithm, callded BSE-VoxNets, is proposed. It achieves superior accuracy and resolution compared to existing methods like MED-SEM and 3D U-Net. By utilizing the Nebula simulator to generate a comprehensive dataset and employing a mixed CNN-transformer backbone architecture, BSE-VoxNets demonstrated its capability to reconstruct core-shell nanocube and nanowire structures with high precision. The results indicate a resolution of 2 nm for both lateral and depth dimensions, outperforming MED-SEM's resolution of 4 nm and 10 nm, respectively. BSE-VoxNets not only advances the field of 3D reconstruction from SEM BSE images but also holds promise for applications in materials science and integrated circuit analysis. Future research will aim to enhance the ability of the model to handle more complex structures and further improve generalization across diverse scenarios.

REFERENCES

[1] M. Holler, M. Guizar-Sicairos, E. H. R. Tsai, R. Dinapoli, E. Müller, O. Bunk, J. Raabe, and G. Aeppli, "High-resolution non-destructive three-dimensional imaging of integrated circuits," *Nature*, vol. 543, no. 7645, pp. 402–406, Mar 2017.

[2] M. Dierolf, A. Menzel, P. Thibault, P. Schneider, C. M. Kewish, R. Wepf, O. Bunk, and F. Pfeiffer, "Ptychographic x-ray computed tomography at the nanoscale," *Nature*, vol. 467, no. 7314, pp. 436–439, Sep 2010.

[3] S. Lee, H. Younan, Z. Siping, and M. Zhiqiang, "Studies on electron penetration versus beam acceleration voltage in energy-dispersive x-ray microanalysis," in *2006 IEEE International Conference on Semiconductor Electronics*, 2006, pp. 610–613.

[4] J. Liu, R. E. Saw, and Y.-H. Kiang, "Calculation of effective penetration depth in x-ray diffraction for pharmaceutical solids," *Journal of Pharmaceutical Sciences*, vol. 99, no. 9, pp. 3807–3814, 2010.

[5] C. O. S. Sorzano, J. Vargas, J. Otón, J. M. de la Rosa-Trevín, J. L. Vilas, M. Kazemi, R. Melero, L. del Caño, J. Cuenca, P. Conesa, J. Gómez-Blanco, R. Marabini, and J. M. Carazo, "A survey of the use of iterative reconstruction algorithms in electron microscopy," *BioMed Research International*, vol. 2017, no. 1, p. 6482567, 2017.

[6] C. Sorzano, J. Vargas, J. de la Rosa-Trevín, A. Jiménez, D. Maluenda, R. Melero, M. Martínez, E. Ramírez-Aportela, P. Conesa, J. Vilas, R. Marabini, and J. Carazo, "A new algorithm for high-resolution reconstruction of single particles by electron microscopy," *Journal of Structural Biology*, vol. 204, no. 2, pp. 329–337, 2018.

[7] L. Bogensperger, E. Kobler, D. Pernitsch, P. Kotzbeck, T. R. Pieber, T. Pock, and D. Kolb, "A joint alignment and reconstruction algorithm for electron tomography to visualize in-depth cell-to-cell interactions," *Histochemistry and Cell Biology*, vol. 157, no. 6, pp. 685–696, Jun 2022.

[8] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "Mlp-mixer: An all-mlp architecture for vision," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 24 261–24 272.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[10] A. Farshian, M. Götz, G. Cavallaro, C. Debus, M. Nießner, J. A. Benediktsson, and A. Streit, "Deep-learning-based 3-d surface reconstruction—a survey," *Proceedings of the IEEE*, vol. 111, no. 11, pp. 1464–1501, 2023.

[11] H. Kniesel, T. Ropinski, T. Bergner, K. S. Devan, C. Read, P. Walther, T. Ritschel, and P. Hermosilla, "Clean implicit 3d structure from noisy 2d stem images," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20 730–20 740.

[12] F. de Isidro-Gómez, J. Vilas, P. Losana, J. Carazo, and C. Sorzano, "A deep learning approach to the automatic detection of alignment errors in cryo-electron tomographic reconstructions," *Journal of Structural Biology*, vol. 216, no. 1, p. 108056, 2024.

[13] T. Houben, T. Huisman, M. Pisarenco, F. van der Sommen, and P. de With, "Training procedure for scanning electron microscope 3D surface reconstruction using unsupervised domain adaptation with simulated data," *Journal of Micro/Nanopatterning, Materials, and Metrology*, vol. 22, no. 3, p. 031208, 2023.

[14] F. Boughorbel, X. Zhuge, P. Potocek, and B. Lich, "SEM 3D Reconstruction of Stained Bulk Samples using Landing Energy Variation and Deconvolution," *Microscopy and Microanalysis*, vol. 18, no. S2, pp. 560–561, 11 2012.

[15] M. de Goede, E. Johlin, B. Sciacca, F. Boughorbel, and E. C. Garnett, "3d multi-energy deconvolution electron microscopy," *Nanoscale*, vol. 9, pp. 684–689, 2017.

[16] E. Kieft and E. Bosch, "Refinement of monte carlo simulations of electron–specimen interaction in low-voltage sem," *Journal of Physics D: Applied Physics*, vol. 41, no. 21, p. 215310, oct 2008.

[17] P. Cizmar, A. E. Vladár, B. Ming, and M. T. a. Postek, "Simulated sem images for resolution measurement," *Scanning*, vol. 30, no. 5, pp. 381–391, 2008.
[18] K. T. Arat, J. Bolten, T. Klimpel, and N. Unal, "Electric fields in Scanning Electron Microscopy simulations," in *Metrology, Inspection, and Process Control for Microlithography XXX*, M. I. Sanchez, Ed., vol. 9778, International Society for Optics and Photonics. SPIE, 2016, p. 97780C.
[19] L. van Kessel and C. Hagen, "Nebula: Monte carlo simulator of electron–matter interaction," *SoftwareX*, vol. 12, p. 100605, 2020.
[20] T. Yu, X. Li, Y. Cai, M. Sun, and P. Li, "S^2-mlpv2: Improved spatial-shift mlp architecture for vision," 2021.
[21] Z. Lu, H. Xie, C. Liu, and Y. Zhang, "Bridging the gap between vision transformers and convolutional neural networks on small datasets," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
[23] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, no. null, p. 1137–1155, Mar. 2003.
[24] M. Berman, A. Rannen Triki, and M. B. Blaschko, "The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4413–4421.
[25] Özgün Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," 2016.
[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.

CAIZHI ZHENG received the B.S. degrees in applied physics from University of Electronic Science and Technology of China, Chengdu, China, in 2022. He is currently pursuing the M.D. degree with the Institute of Electromagnetics and Acoustics, Xiamen University, Xiamen, China. His research interests include computer vision, deep learning, and 3D reconstruction problems.

RONGHAN HONG (Member, IEEE) received the B.S degree in electronic information science and technology and the Ph.D. degree in radio physics from Xiamen University, Xiamen, China, in 2014 and 2021, respectively. His current research interests include computational electromagnetics and inverse problem for 3D reconstruction.

HAO-JIE HU received the B.S. degree in communication engineering from Zhengzhou University, Zhengzhou, China, in 2019. He is currently pursuing the Ph.D. degree with Xiamen University, Xiamen, China. His research interests include electromagnetic inverse problems and deep neural network techniques.

QING HUO LIU (Fellow, IEEE) received the B.S. and M.S. degrees in physics from Xiamen University, Xiamen, China, in 1983 and 1986, respectively, and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 1989. He was a Post-Doctoral Research Associate with the University of Illinois at Urbana-Champaign in 1990 and the Research Scientist and the Program Leader with Schlumberger-Doll Research, Ridge field, CT, USA, from 1990 to 1995. From 1996 to May 1999, he was an Associate Professor with New Mexico State University, Las Cruces, NM, USA. Since June 1999, he has been with Duke University, Durham, NC, USA, where he is currently a Professor of electrical and computer engineering. He has been also the Founder and the Chairman of Wave Computation Technologies, Inc., Durham, since 2005. From 2019 to 2020, he was a Visiting Professor with Kyoto University, Kyoto, Japan. Since 2011, he has been a Visiting Professor with Xiamen University. His research interests include computational electromagnetics and acoustics, inverse prob lems, and their application in nanophotonics, geophysics, biomedical imaging, and electronic design automation. He has published widely in these areas. Dr. Liu is a Fellow of the Acoustical Society of America, the Elec tromagnetics Academy, and the Optical Society of America. He received the 1996 Presidential Early Career Award for Scientists and Engineers (PECASE) from the White House, the 1996 Early Career Research Award from the Environmental Protection Agency, the 1997 CAREER Award from the National Science Foundation, the 2017 Technical Achievement Award, the 2018 Computational Electromagnetics Award from the Applied Com putational Electromagnetics Society, the 2018 Harrington-Mittra Award in Computational Electromagnetics from the IEEE Antennas and Propagation Society, and the ECE Distinguished Alumni Award from the University of Illinois at Urbana-Champaign in 2018. He has served as an IEEE Antennas and Propagation Society Distinguished Lecturer, and the Founding Editor-in Chief for the *IEEE Journal on Multiscale and Multiphysics Computational Techniques.*