Divide and Conquer Strategy on Monocular Depth Estimation of Semiconductor SEM Images

Hao Sun*, Peijin Cai†, Sihai Zhang*‡
*School of Integrated Circuits, University of Science and Technology of China, Hefei, Anhui, P.R. China
†Institute of Advanced Technology, University of Science and Technology of China, Hefei, Anhui, P.R. China
‡Anhui Province Key Laboratory of Integrated Circuit Science and Technology
Email: {sun0405,peijinc}@mail.ustc.edu.cn, shzhang@ustc.edu.cn

***Abstract*—Single-image scanning electron microscope (SEM) depth estimation plays a critical role in semiconductor three-dimensional metrology. However, existing end-to-end approaches often struggle to maintain stable generalization performance in complex structural scenarios composed of multiple geometric primitives. To address this issue, we propose a Divide and Conquer style structured modeling approach for single-image SEM depth estimation which explicitly decomposes complex scenes into multiple local structural regions and shares depth prediction parameters across regions, thereby guiding the model to learn reusable local structure–depth mapping relationships and reducing its reliance on global layout statistics. The proposed approach is systematically validated on a simulation-generated SEM dataset. The training data include both basic structural primitives and their compositions forming SRAM-like structures, while the testing phase further evaluates the model's prediction stability under variations in structural compositions. Experimental results demonstrate that, compared with conventional end-to-end models, our method achieves more stable and reliable depth estimation performance in complex compositional scenarios, validating the effectiveness of structured modeling strategies in improving the generalization capability of single-image SEM depth estimation.**

***Index Terms*—Machine learning, monocular depth estimation, scanning electron microscopy, divide-and-conquer strategy.**

I. Introduction

Recent advances in model scale and architectural complexity have substantially improved visual learning across many tasks [1], [2]. Yet when training data do not adequately cover structural compositional variations in real applications, holistic end-to-end models often generalize poorly. To address this issue, recent studies have explored explicit modeling of local structures and their compositional relationships, so that learned local patterns can be reused under unseen combinations [3].

Compared with multi-view imaging or hardware-modification-based solutions [4], single-image SEM depth estimation offers advantages in reducing imaging cost and system complexity, and has therefore attracted sustained attention. Existing studies are mainly conducted on SEM datasets with relatively simple structures or well-controlled imaging conditions [5], demonstrating the feasibility of deep learning-based approaches. However, in highly integrated chip layouts with complex structural compositions, their prediction performance still faces significant challenges.

Most existing single-image SEM depth estimation methods adopt an end-to-end framework [6] that directly maps grayscale images to depth values. Such models tend to entangle local geometric cues with global layout information. In highly integrated structures such as SRAM, SEM intensity is jointly affected by local geometry and overall layout, which can lead to degraded performance under unseen structural compositions. This reveals a compositional generalization problem that remains insufficiently studied in current SEM depth estimation research.

Beyond prediction accuracy, poor compositional generalization also limits the practical deployment of SEM depth estimation in semiconductor metrology. The same local structures often appear under different neighboring contexts, process variations, and imaging conditions. Models that rely too heavily on global layout statistics may work well on familiar layouts but become unstable under new compositions. This practical challenge motivates a modeling strategy that separates reusable local geometric patterns from scene-level compositional relationships.

To address this problem, we propose a structured compositional framework for monocular SEM depth estimation. The method explicitly decomposes a complex scene into multiple local structural regions and applies shared depth prediction across them, encouraging the learning of reusable local structure–depth mappings. This design preserves end-to-end training while reducing reliance on global layout statistics, thereby improving robustness under structural compositional variations.

The contributions of this paper are as follows:

- We propose a structured framework for single-image SEM depth estimation that combines explicit structural decomposition with shared depth prediction, enabling the reuse of local structural primitives and their depth mappings for more stable prediction under complex structural compositions.
- We validate the proposed method on a simulation-generated SEM dataset containing both basic structures and SRAM-like complex compositions, and show that it achieves more stable depth estimation performance than conventional end-to-end models in complex structural scenarios.

II. Related Work

A. SEM Image Depth Estimation

Early studies on SEM image depth estimation mainly relied on specific imaging conditions or dedicated hardware configurations. Representative approaches include three-dimensional reconstruction from stereo image pairs acquired by sample tilting [7], as well as surface profile recovery by combining intensity information from multi-channel secondary electron detectors [8]. These methods can provide useful geometric cues, but they usually depend on dedicated SEM systems or additional acquisition procedures, which limits their flexibility in practical complex scenarios.

Beyond hardware-based solutions, some studies attempt to estimate depth by analyzing SEM imaging signals or incorporating physical models. For example, observed SEM images can be matched against pre-established physical model libraries [9], or height information can be inferred from top-down SEM images by exploiting the relationship between incident electron beam energy and signal intensity [10]. Although these methods improve interpretability to some extent, they often rely on known structural models or specific imaging settings, which differ from the low-energy SEM configurations commonly used in semiconductor manufacturing.

In recent years, deep learning-based approaches for single-image SEM depth estimation have attracted increasing attention. These methods typically adopt end-to-end learning frameworks that directly regress pixel-wise depth information from two-dimensional SEM grayscale images, and have shown promising performance under relatively simple structural settings or well-controlled imaging conditions [11], [12]. However, integrated circuit layouts usually contain highly repetitive local structures together with complex compositional patterns. Under such conditions, end-to-end models tend to entangle local geometric cues with global layout statistics during training, which often leads to degraded generalization when structural compositions vary. This suggests that complex SEM depth estimation requires a more structured modeling strategy than holistic regression alone.

B. Learning with Explicit Structural Decomposition

In complex visual scenes, holistic end-to-end modeling often struggles to capture the compositional relationships among local structures. To alleviate this limitation, prior work has explored learning paradigms based on explicit structural decomposition, where input representations are partitioned into multiple constituent units with relatively independent semantic or geometric meanings, enabling the learning of more reusable intermediate representations [13].

Representative approaches include object-centric learning frameworks such as Slot Attention [14], which use learnable attention mechanisms to decompose a scene into a set of latent structural units and have demonstrated strong structural modeling capability in various visual perception and reasoning tasks. More broadly, these methods often combine region partitioning, attention mechanisms, or latent structure modeling to group or decompose input features, thereby alleviating the dependence of holistic modeling on global statistical regularities [15].

In a broader line of research, explicit structural decomposition has also been regarded as an effective way to improve compositional generalization. Prior studies suggest that when models learn representations based on structural primitives or local units, they are more likely to maintain stable performance under unseen compositional configurations [3], [16]. However, most existing efforts focus on high-level tasks such as semantic object discovery, vision–language reasoning, or generative modeling, where structural decomposition mainly serves abstract semantic or symbolic compositional reasoning.

For dense prediction problems, especially in microscopic imaging scenarios with highly repetitive local structures and complex composition patterns, related modeling strategies remain relatively limited. Overall, existing studies indicate that explicit structural decomposition can enhance structured representation learning, yet its systematic use in dense depth prediction is still underexplored. This gap motivates us to introduce structural decomposition into monocular SEM depth estimation for better robustness under compositional variation.

III. Compositional Generalization Model

This paper proposes a compositional framework for monocular SEM image depth estimation [17]. By introducing structural decomposition and shared depth prediction [18], the model represents complex SEM scenes as combinations of local structural regions, enabling the learning of reusable local primitives and their depth mappings for improved robustness under structural compositional variations.

As shown in Fig. 1, given an input SEM image $I_{\text{sem}} \in \mathbb{R}^{H\times W}$, the model first extracts shared features using a convolutional backbone with a ResNet-style encoder and a U-Net-like decoder, which balances semantic abstraction and spatial detail preservation for dense prediction [19]. Based on the shared features [17], a structural decomposition module partitions the scene into multiple local structural regions. A parameter-shared depth predictor is then applied to each region [18], and the resulting local predictions are fused to reconstruct the final pixel-wise depth map. This design remains end-to-end trainable while reducing reliance on global layout statistics, leading to more stable depth estimation under varying structural compositions.

A. Structural Decomposition

For SEM scenes in complex chip layouts composed of multiple basic structural primitives, we explicitly introduce a structural decomposition mechanism in the feature space. The goal is to represent the overall scene as a compositional combination of several local structural regions with relatively consistent geometric characteristics. It should be noted that this decomposition does not correspond to precise semantic or physical labeling of structures, but serves as a modeling strategy to encourage the learning of reusable local structural representations.

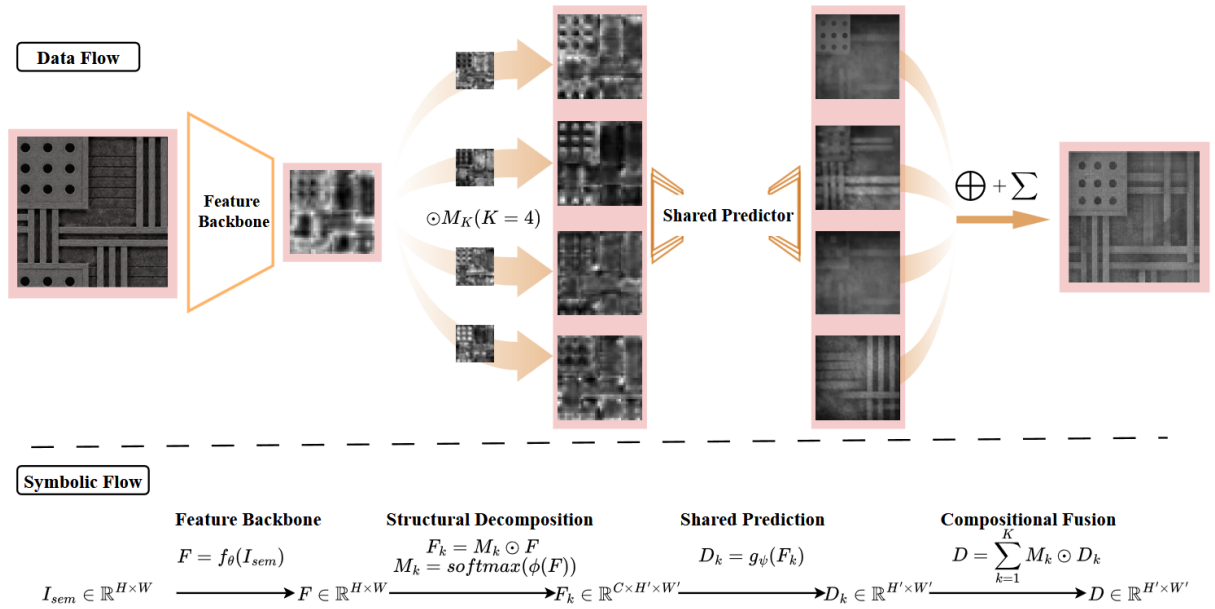

Fig. 1. Overview of the proposed framework. Shared features extracted from the input SEM image are decomposed into K soft masks for local depth prediction, and the final depth map is obtained by weighted fusion. The lower part shows the symbolic formulation.

Given an input SEM image $I_{\text{sem}} \in \mathbb{R}^{H\times W}$, a shared feature map is first extracted by the backbone network:

$$\mathbf{F} = f_\theta(I_{\text{sem}}), \quad \mathbf{F} \in \mathbb{R}^{C\times H'\times W'}, \tag{1}$$

where $f_\theta(\cdot)$ denotes the feature extraction network, C is the number of feature channels, and H', W' are the spatial resolutions after downsampling.

Based on the shared feature representation $\mathbf{F}$, a structural decomposition head is applied to predict K soft structural masks:

$$\{M_k\}_{k=1}^{K} = \text{Softmax}(\phi(\mathbf{F})), \tag{2}$$

where $\phi(\cdot)$ denotes a lightweight convolutional projection applied to $\mathbf{F}$, and each mask $M_k \in [0,1]^{H'\times W'}$ represents the response strength of the k-th structural component at each spatial location.

The softmax normalization is performed across the K components at each spatial position, such that

$$\sum_{k=1}^{K} M_k(i,j) = 1, \quad \forall (i,j), \tag{3}$$

ensuring that the feature response at each location is softly assigned to different structural regions. Through this decomposition, complex SEM scenes are represented as weighted combinations of local structural regions in the feature space, providing structured input for subsequent region-wise depth prediction with shared parameters.

B. Shared Depth Prediction

After structural decomposition, the input SEM image is represented in the feature space as K local structural components [20], each associated with a soft structural mask M_k while sharing the same feature representation F. To prevent the model from learning region-specific depth mapping functions that hinder the reuse of structural primitives, we adopt a *parameter-shared depth prediction strategy* across all structural components.

Specifically, as illustrated in the symbolic flow, the k-th structural component first applies its corresponding soft mask to the shared feature map to obtain a structure-aware regional representation:

$$F_k = M_k \odot F, \tag{4}$$

where $\odot$ denotes element-wise multiplication. This operation encourages each regional feature to focus on its corresponding structural response while maintaining a unified representation form.

All regional features $\{F_k\}_{k=1}^{K}$ are then processed by a single parameter-shared depth prediction head $g_\psi(\cdot)$ to produce region-wise depth predictions:

$$D_k = g_\psi(F_k), \quad k = 1,\ldots,K, \tag{5}$$

where $D_k \in \mathbb{R}^{H'\times W'}$ denotes the depth response of the k-th structural component.

By sharing prediction parameters across different structural regions, the model is constrained to reuse the same depth mapping function under varying spatial locations and structural compositions, rather than overfitting to region-specific layout statistics [21]. This shared prediction mechanism introduces cross-region consistency at the modeling level, endowing the region-wise depth estimation with improved transferability and providing a stable foundation for subsequent compositional fusion.

C. Compositional Fusion

After obtaining region-level depth predictions for all local structural components, the model aggregates these predictions to reconstruct the final global depth map. As illustrated in the symbolic flow, we adopt an explicit mask-weighted compositional fusion mechanism to combine region-wise depth responses, enabling structured reconstruction of the overall depth estimation.

Specifically, the final depth prediction is computed as a pixel-wise weighted summation of the region-level depth maps:

$$D = \sum_{k=1}^{K} M_k \odot D_k, \tag{6}$$

where $D \in \mathbb{R}^{H' \times W'}$ denotes the fused depth map, M_k is the soft structural mask corresponding to the k-th component, and $\odot$ represents element-wise multiplication. Since the structural masks are normalized via a softmax operation at each spatial location during the decomposition stage, this fusion process can be interpreted as a *compositional reconstruction*, in which the global depth is formed as a linear combination of local structural depth responses.

Through this compositional fusion strategy, the proposed compositional depth estimation framework is explicitly constrained to predict the global depth map as an aggregation of region-level structural predictions, rather than directly regressing depth from holistic feature representations. This design reinforces the role of local structural modeling in global depth estimation and further contributes to stable predictions under varying structural compositions.

IV. Experiments

In this section, we describe the experimental setup and evaluation protocol used to assess the effectiveness of the proposed compositional modeling approach for single-image SEM depth estimation.

A. Datasets

The dataset is generated using the Nebula SEM imaging simulator [22]. Each sample contains a single SEM image and its pixel-wise depth ground truth. The simulator models key imaging factors, including electron beam acceleration voltage, detector response, and material parameters, so that the synthesized images remain consistent with real SEM characteristics in grayscale distribution and structural response.

The dataset contains four structure categories, denoted as S1–S4. S1–S3 correspond to three basic structures: contact holes, line/space, and trench. Each basic category is split into training and testing sets with an 80% / 20% ratio. During data generation, both geometric parameters, such as hole diameter, line width, and trench size, and imaging conditions, such as beam energy and depth range, are randomized to cover the parameter space of local structural primitives while keeping the global layout relatively simple.

In addition, we construct S4, an SRAM-like compositional dataset. Its samples are generated by combining the primitives in S1–S3 with more complex spatial arrangements to mimic typical integrated circuit layouts. S4 is also divided into training and testing sets with an 80% / 20% ratio. Although SRAM-like structures are included in training, the test samples use different spatial arrangements and layout patterns, enabling evaluation of generalization under structural compositional variations.

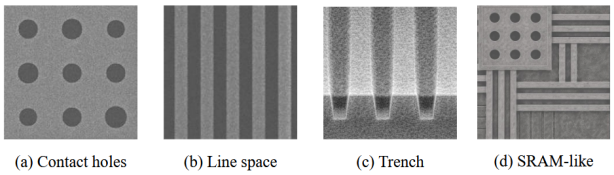

Fig. 2. Representative SEM images in the dataset: (a) contact holes, (b) line/space, (c) trench, and (d) an SRAM-like compositional structure.

The sample counts and corresponding geometric and imaging settings are summarized in Table I. Representative SEM images are shown in Fig. 2.

B. Experimental Setting

The proposed model is implemented in PyTorch, and all experiments are conducted on a workstation equipped with an NVIDIA GeForce RTX 3090 GPU with 24 GB memory. Training uses the AdamW optimizer with an initial learning rate of 2×10^{-4}. The learning rate of the feature backbone is set to $0.25\times$ that of the prediction modules to maintain stable shared feature representations during end-to-end training. A step-wise learning rate decay strategy is adopted, and the learning rate is reduced once the validation performance converges. During training, input SEM images are cropped into 256×256 patches, with a batch size of 8.

We systematically investigate the impact of the number of structural decomposition components K on model performance, with $K = 1, 2, 4,$ and 8 evaluated for comparison. Among them, $K = 1$ corresponds to an end-to-end setting without explicit structural decomposition and serves as the baseline within the proposed framework. Except for the value of K, all other network architectures and training configurations are kept identical across experiments.

Model training is performed on the simulation-generated SEM dataset using a staged end-to-end optimization strategy. The model is first pre-trained on the training data of S1–S3 to stabilize the mask prediction behavior of the structural decomposition module and to guide the shared depth prediction module to learn consistent local structure–depth mapping relationships. Training samples including S4 are then introduced for joint training, allowing the structural decomposition, shared prediction, and compositional reconstruction modules to co-converge within an end-to-end framework. During this stage, the SRAM-like structures in the test set do not overlap with the training samples in specific spatial arrangements, thereby preventing data leakage at the level of structural composition.

In addition, the proposed method is compared with representative single-image SEM depth estimation approaches,

TABLE I
GEOMETRIC PARAMETER AND IMAGING CONDITION CONFIGURATIONS FOR THE BASIC AND COMPOSITIONAL STRUCTURES USED IN THE TRAINING DATASET.

Dataset ID	Structure Type	Amount(piece)	Beam Energy (eV)	Depth Range (nm)	Geometry Parameters(nm)
S1	Contact holes	5000	1000,1500	200-300	Diameter: 20–30
S2	Line space	5000	1000,1500	200-300	Line width: 20–30
S3	Trench	5000	1000,1500	200-300	Step height: 18–32
S4	SRAM-like	2000	1000,1500	200-300	Composed from basic primitives

including Eigen [23], BTS [24], and Houben. All comparison methods are evaluated under the same data splits and training/testing settings to ensure fairness and comparability. Performance is assessed using two standard depth estimation metrics: MAE and RMSE.

C. Results and Discussions

Table II summarizes the quantitative results on the S1–S4 test sets. Overall, the proposed method outperforms representative single-image SEM depth estimation approaches, including Eigen, BTS, and Houben, on both simple structural primitives and SRAM-like compositional structures. The advantage becomes more pronounced on the compositional test set, indicating better robustness and generalization under structural compositional variations.

Among all settings, the four-component setting achieves the best overall performance. Compared with Houben, using four decomposition components reduces the MAE by 40.87% and the RMSE by 38.13% on the contact-hole test set. The gain is even larger on SRAM-like compositional structures, where the MAE and RMSE are reduced by 66.25% and 43.70%, respectively. This suggests that four components provide the most suitable decomposition granularity for the SEM depth estimation task in this study. As shown in Fig. 3, this setting yields clearer responses to different structural regions, indicating a better balance between structural expressiveness and prediction stability. The complex SEM scenes in our dataset are mainly composed of contact holes, line structures, and trench structures, while one additional component can account for transition or mixed regions. Therefore, the four-component setting can cover the dominant structural patterns while retaining sufficient flexibility, leading to the best overall performance.

When $K < 4$, the model does not reach its best performance mainly because the decomposition granularity is insufficient. For $K = 1$, the model essentially degenerates into holistic end-to-end regression without explicit structural decomposition, and is therefore more likely to rely on global layout statistics. For $K = 2$, although structural decomposition is introduced and the performance improves over $K = 1$, two components are still insufficient to separate the major structural primitives effectively. As a result, some local structures remain mixed within the same component, and the shared predictor still receives entangled region-wise features. This limits the stable learning of reusable local structure–depth mappings. In essence, the limitation of small K lies in under-decomposition, where the

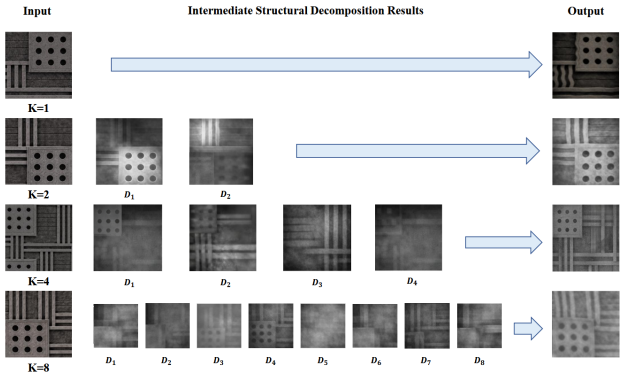

Fig. 3. Qualitative results with different numbers of components K.

structural partition is not fine enough to explicitly characterize the main local patterns in complex SEM scenes.

When $K > 4$, the performance decreases again, but for a different reason. In this case, the issue is no longer insufficient decomposition but over-decomposition. Taking $K = 8$ as an example, Fig. 3 shows that the structural masks become fragmented and the corresponding depth responses are more dispersed. This indicates that an excessively large K tends to split originally coherent structural regions into overly fine fragments, which weakens the consistency of region-wise representations and makes it harder for the shared predictor to learn stable local structure–depth mappings. Therefore, the limitation of large K lies in over-decomposition: adding more components does not improve structural modeling, but instead harms the integrity of local structural representations.

Taken together, these results show that the improvement in complex compositional scenes mainly comes from an appropriate decomposition granularity rather than from increased model capacity alone. A smaller K leads to under-decomposition, whereas a larger K leads to over-decomposition. In this study, $K = 4$ provides a favorable balance between the two, enabling the model to reuse local structural primitives more reliably and to achieve more stable depth estimation in complex SEM scenes.

V. CONCLUSIONS

This paper addresses the generalization degradation of single-image SEM depth estimation under complex structural compositions by proposing a structured modeling framework based on structural decomposition and shared depth prediction. By decomposing complex SEM scenes into reusable local structural units and sharing prediction parameters across

TABLE II

QUANTITATIVE COMPARISON OF DEPTH ESTIMATION PERFORMANCE UNDER DIFFERENT STRUCTURAL COMPLEXITIES. MODELS ARE EVALUATED ON SIMPLE STRUCTURAL PRIMITIVES (CONTACT HOLE AND TRENCH) AND SRAM-LIKE COMPOSITIONAL STRUCTURES. RELATIVE PERFORMANCE CHANGES WITH RESPECT TO THE HOUBEN METHOD ARE REPORTED IN PARENTHESES.

Method	**Simple Structures**				**Compositional Structures**	
	Contact Hole		Trench			
	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
Eigen	32.845(+73.35%)	53.489(+77.03%)	28.764(+93.54%)	44.931(+91.22%)	90.464(+48.68%)	150.467(+66.29%)
BTS	22.316(+17.78%)	35.742(+18.29%)	19.904(+33.93%)	29.856(+27.06%)	68.489(+12.56%)	100.924(+11.53%)
Houben	18.947(0)	30.215(0)	14.862(0)	23.497(0)	60.846(0)	90.487(0)
Ours (K=1)	14.576(-23.07%)	23.478(-22.30%)	10.784(-27.44%)	16.487(-29.83%)	40.464(-33.50%)	70.965(-21.57%)
Ours (K=2)	13.234(-30.15%)	22.014(-27.14%)	9.987(-32.80%)	15.863(-32.49%)	31.943(-47.50%)	62.478(-30.95%)
Ours (K=4)	**11.203(-40.87%)**	**18.694(-38.13%)**	**8.216(-44.72%)**	**13.052(-44.45%)**	**20.534(-66.25%)**	**50.948(-43.70%)**
Ours (K=8)	15.179(-19.89%)	24.015(-20.52%)	10.842(-27.05%)	16.978(-27.74%)	36.714(-39.66%)	66.964(-26.00%)

regions, the model learns more stable local structure–depth mappings. Experimental results on SRAM-like compositional structures show that the proposed method achieves more stable depth estimation than conventional end-to-end models, demonstrating the effectiveness of structure-level modeling for improving generalization in single-image SEM depth estimation. Future work may include adaptive structural decomposition and evaluate the method on more complex real SEM scenarios.

ACKNOWLEDGMENT

This work was partially supported by the Anhui Provincial Science and Technology Breakthrough Project under Grant No.202523o09050015 and Advanced Micro-Fabrication Equipment Inc. China (AMEC) Innovative Research under Grant No. EF2190000055.

REFERENCES

[1] M. Rodrigo, C. Cuevas, and N. García, "Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks," *Scientific reports*, vol. 14, no. 1, p. 21392, 2024.

[2] A. Mertan, D. J. Duff, and G. Unal, "Single image depth estimation: An overview," *Digital Signal Processing*, vol. 123, p. 103441, 2022.

[3] K. Farid, R. Sahay, Y. A. Alnaggar, S. Schrodi, V. Fischer, C. Schmid, and T. Brox, "What drives compositional generalization in visual generative models?" *arXiv preprint arXiv:2510.03075*, 2025.

[4] A. Baghaie, A. Pahlavan Tafti, H. A. Owen, R. M. D'Souza, and Z. Yu, "Three-dimensional reconstruction of highly complex microscopic samples using scanning electron microscopy and optical flow estimation," *PloS one*, vol. 12, no. 4, p. e0175078, 2017.

[5] T. Houben, T. Huisman, M. Pisarenco, F. Van Der Sommen, and P. H. de With, "Depth estimation from a single sem image using pixel-wise fine-tuning with multimodal data," *Machine Vision and Applications*, vol. 33, no. 4, p. 56, 2022.

[6] T. Houben, M. Pisarenco, T. Huisman, H. Onvlee, F. van der Sommen *et al.*, "Depth estimation from sem images using deep learning and angular data diversity," in *Metrology, Inspection, and Process Control XXXVII*, vol. 12496. SPIE, 2023, pp. 379–387.

[7] S. Roy, J. Meunier, A. M. Marian, F. Vidal, I. Brunette, and S. Costantino, "Automatic 3d reconstruction of quasi-planar stereo scanning electron microscopy (sem) images," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 4361–4364.

[8] W. Ito, B. Bunday, S. Harada, A. Cordes, T. Murakawa, A. Arceo, M. Yoshikawa, T. Hara, T. Arai, S. Shida *et al.*, "Novel three dimensional (3D) cd-sem profile measurements," in *Metrology, Inspection, and Process Control for Microlithography XXVIII*, vol. 9050. SPIE, 2014, pp. 85–93.

[9] A. E. Vladár, J. S. Villarrubia, J. Chawla, B. Ming, J. R. Kline, S. List, and M. T. Postek, "10nm three-dimensional cd-sem metrology," in *Metrology, Inspection, and Process Control for Microlithography XXVIII*, vol. 9050. SPIE, 2014, pp. 53–63.

[10] K. T. Arat, J. Bolten, A. C. Zonnevylle, P. Kruit, and C. W. Hagen, "Estimating step heights from top-down sem images," *Microscopy and Microanalysis*, vol. 25, no. 4, pp. 903–911, 2019.

[11] J. Park, Y. Cho, Y. Hwang, A. Ma, Q. Kim, K.-B. Chang, J. Jeong, and S.-J. Kang, "Mixup-based neural network for image restoration and structure prediction from sem images," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[12] Y. Hwang, M. Song, A. Ma, Q. Kim, K.-B. Chang, J. Jeong, and S.-J. Kang, "So-diffusion: Diffusion-based depth estimation from sem images and ocd spectra," *IEEE Transactions on Instrumentation and Measurement*, 2025.

[13] C. P. Burgess, L. Matthey, N. Watters, R. Kabra, I. Higgins, M. Botvinick, and A. Lerchner, "Monet: Unsupervised scene decomposition and representation," *arXiv preprint arXiv:1901.11390*, 2019.

[14] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," *Advances in neural information processing systems*, vol. 33, pp. 11 525–11 538, 2020.

[15] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," *Computational visual media*, vol. 9, no. 4, pp. 733–752, 2023.

[16] C. Li, Z. Li, C. Jing, Y. Jia, and Y. Wu, "Exploring the effect of primitives for compositional generalization in vision-and-language," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 092–19 101.

[17] T. Monnier, J. Austin, A. Kanazawa, A. Efros, and M. Aubry, "Differentiable blocks world: Qualitative 3d decomposition by rendering primitives," *Advances in Neural Information Processing Systems*, vol. 36, pp. 5791–5807, 2023.

[18] J. Huang, J. Gao, V. Ganapathi-Subramanian, H. Su, Y. Liu, C. Tang, and L. J. Guibas, "Deepprimitive: Image decomposition by layered primitive detection," *Computational Visual Media*, vol. 4, no. 4, pp. 385–397, 2018.

[19] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *International workshop on deep learning in medical image analysis*. Springer, 2016, pp. 179–187.

[20] P.-Y. Chen, A. H. Liu, Y.-C. Liu, and Y.-C. F. Wang, "Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2019, pp. 2624–2632.

[21] L. Wang, J. Zhang, O. Wang, Z. Lin, and H. Lu, "Sdc-depth: Semantic divide-and-conquer network for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 541–550.

[22] L. van Kessel and C. Hagen, "Nebula: Monte carlo simulator of electron–matter interaction," *SoftwareX*, vol. 12, p. 100605, 2020.

[23] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in neural information processing systems*, vol. 27, 2014.

[24] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *arXiv preprint arXiv:1907.10326*, 2019.