End-to-End 3D Reconstruction Network with Occlusion Detectors for Multi-View FPP

Zuqiong Chen
Mechatronics and Control Engineering
Shenzhen University
Shenzhen, China
2410094001@mails.szu.edu.cn

Ruisheng Wang
Architecture & Urban Planning
Shenzhen University
Shenzhen, China
ruiswang@szu.edu.cn

Yibin Tian
Mechatronics and Control Engineering
Shenzhen University
Shenzhen, China
ybtian@szu.edu.cn

Abstract*—Most existing Deep learning (DL) –based 3D reconstruction methods for Fringe Projection Profilometry (FPP) rely on single-view inputs and suffer from occlusions and incomplete reconstruction when applied to complex geometries. To address these limitations, we propose an end-to-end DL framework for high-fidelity 3D reconstruction from fringe patterns acquired from four complementary views. The proposed network adopts a multi-task architecture built upon a shared 3D residual encoder–decoder backbone, augmented with an Atrous Spatial Pyramid Pooling (ASPP) module to capture multi-scale contextual information without sacrificing spatial resolution. To explicitly handle view-dependent occlusions arising from complex objects, a lightweight occlusion detector and an adaptive occlusion-aware fusion module are introduced to guide multi-view feature aggregation. In parallel, a self-supervised view reconstruction branch is jointly optimized with depth estimation, acting as a structural regularizer that promotes geometric consistency and preserves fine details in the latent feature space. Extensive experiments conducted on a large-scale synthetic dataset demonstrate that the proposed method achieves a Mean Absolute Error (MAE) of 0.0447 mm and a Root Mean Square Error (RMSE) of 0.1741 mm, representing a more than 50% reduction in reconstruction error compared with state-of-the-art single-view DL approaches. These results establish a new benchmark for robustness and completeness in FPP 3D measurement under challenging occlusion and complex geometry conditions.

***Keywords*—Fringe projection profilometry, Occlusion, 3D reconstruction, Deep neural network, Multi-task learning**

I. INTRODUCTION

Fringe Projection Profilometry (FPP) has been widely adopted for high-accuracy 3D shape measurement, owing to its full-field acquisition capability, high spatial resolution, and flexibility in both laboratory and industrial environments. The principle of FPP involves projecting a sequence of fringe patterns onto a target surface using a digital projector, while one or more cameras capture the deformed fringe images. The object's surface geometry modulates the projected fringes, and the resulting phase variations are subsequently decoded to recover dense depth information with sub-pixel precision [1-6]. Compared with point-based or scanning methods, FPP offers superior measurement throughput and is particularly well suited for rapid inspection of complex surfaces.

In recent years, the rapid development of deep neural networks (DNNs) has reshaped the FPP processing pipeline. Learning-based approaches have demonstrated strong robustness against noise, defocus, low contrast, and non-ideal illumination conditions. DNNs have been successfully applied to a broad range of FPP-related tasks, including fringe pattern analysis and phase extraction [7-8], fringe denoising and enhancement [9-10], phase unwrapping under discontinuous or high-gradient surfaces [11-12], as well as end-to-end depth regression that directly maps fringe images to 3D geometries [13-16]. These data-driven models not only reduce the reliance on handcrafted algorithms and strict system calibration, but also enable real-time or near real-time reconstruction in practical deployment scenarios.

Despite the advances, most state-of-the-art FPP reconstruction DNNs are for the monocular paradigm, assuming a single camera–projector configuration. Monocular FPP suffers from visibility constraints imposed by the imaging geometry. Occlusions, self-shadowing, and steep surface discontinuities can lead to blind spots and incomplete 3D reconstruction. This limitation becomes particularly pronounced in industrial inspections involving complex geometries, such as printed circuit board (PCB), where densely populated components, tall structures, and sharp edges frequently obstruct the projected fringes or the camera's view. Consequently, monocular DNN-based FPP often struggles to achieve complete and reliable 3D coverage.

To alleviate the limitations inherent to monocular FPP, multi-view imaging systems have been extensively explored, in which multiple cameras and/or projectors are distributed around the object to ensure that each surface patch is observed by at least one sensor [17-20]. By exploiting complementary viewpoints, multi-view FPP systems can significantly improve surface coverage, reduce missing regions, and enhance reconstruction completeness for objects with high aspect-ratio structures or depth discontinuities.

Conventional multi-view FPP pipelines are burdened by a series of nontrivial challenges. These methods typically require accurate geometric calibration of all cameras and projectors, followed by explicit point cloud registration or depth map fusion across views. Such processes are highly sensitive to calibration errors and mechanical misalignment, which are common in industrial environments. Even minor deviations in intrinsic or extrinsic parameters can propagate through the reconstruction pipeline, leading to accumulated errors, surface distortion, or stitching artifacts. Furthermore, classical registration algorithms—such as those based on Iterative Closest Point (ICP) or feature correspondence—often suffer from local minima, degraded performance on textureless or repetitive structures, and increased computational cost.

End-to-end DNN approaches provide an alternative by reframing multi-view fusion as a unified learning and optimization problem. These networks can directly ingest multi-view fringe images or intermediate representations and learn to aggregate complementary information in a data-driven manner. By leveraging shared latent spaces, attention mechanisms, or geometric consistency constraints, such models are capable of implicitly learning inter-view relationships, visibility priors, and cross-view correspondences. As a result, global surface consistency can be enforced during inference without explicit point cloud

registration or iterative alignment, leading to improved robustness against calibration inaccuracies and environmental perturbations.

In this study, we propose an end-to-end multi-task learning framework, termed ODMV-Net, which directly processes structured-light measurements acquired from multiple viewpoints to achieve high-precision depth reconstruction. Unlike conventional pipelines that decouple view-specific reconstruction and geometric fusion, ODMV-Net jointly optimizes multi-view feature extraction, occlusion handling, and depth inference within a unified network. It incorporates a self-supervised view reconstruction branch that serves as a structural regularizer, encouraging the latent 3D feature representation to preserve fine-grained geometric details and cross-view consistency. The architecture is built upon a hierarchical 3D residual encoder-decoder backbone, augmented with Atrous Spatial Pyramid Pooling (ASPP) [21-24] to effectively capture multi-scale contextual dependencies while maintaining spatial resolution. This design enables the network to model both local surface details and global geometric structures across views. To explicitly address view-dependent occlusions, we introduce a lightweight OCclusion Detector (OCD) alongside an Occlusion-Aware Fusion (OAF) module. The OCD predicts per-view occlusion masks that identify unreliable or missing regions, while the OAF module leverages these masks to dynamically modulate skip connections and guide multi-view feature fusion during decoding. By suppressing occluded responses through learned attention weights, the proposed modules enable adaptive integration of complementary information from visible regions across viewpoints.

Through the synergistic interaction of self-supervised regularization, multi-scale 3D feature encoding, and occlusion-aware fusion, ODMV-Net achieves robust and accurate depth reconstruction even in challenging scenes characterized by severe occlusion and complex geometry.

II. Method

A. ODMV-Net Architecture

The ODMV-Net is an end-to-end multi-task learning network for high-precision depth reconstruction from multi-view fringe pattern measurements. An overview of the network is shown in Fig. 1. Built upon a unified 3D encoder–decoder backbone with hard parameter sharing, ODMV-Net jointly processes structured-light inputs from multiple viewpoints, enabling consistent feature extraction and cross-view interaction within a single optimization framework. The backbone comprises four hierarchical encoding stages, each constructed from stacked 3D residual blocks, which progressively capture geometric features from low-level fringe modulations to high-level structural representations.

Preceding the backbone, a Residual View Encoder, shown in Fig. 2(a), is employed to initialize multi-view feature representation. Specifically, spatial features from each viewpoint are first extracted using independent strided 2D convolutions and subsequently stacked to form a unified 3D feature volume. Inter-view dependencies are then explicitly modeled through anisotropic convolutions that operate along the view dimension, enabling effective cross-view interaction. These view-aware features are further refined via residual connections to facilitate stable optimization and preserve view-specific information.

At the bottleneck of the network, a 3D ASPP module is introduced to capture rich multi-scale contextual information by aggregating features with varying receptive fields, while maintaining spatial resolution. In parallel, a lightweight OCD, as shown in Fig. 2(b), is designed to identify view-dependent valid and invalid regions. It leverages efficient depthwise separable convolutions to predict soft occlusion masks (ODmask), which encode the reliability of features across spatial locations. During decoding, these masks are used to actively modulate skip connections, selectively suppressing occluded or unreliable features.

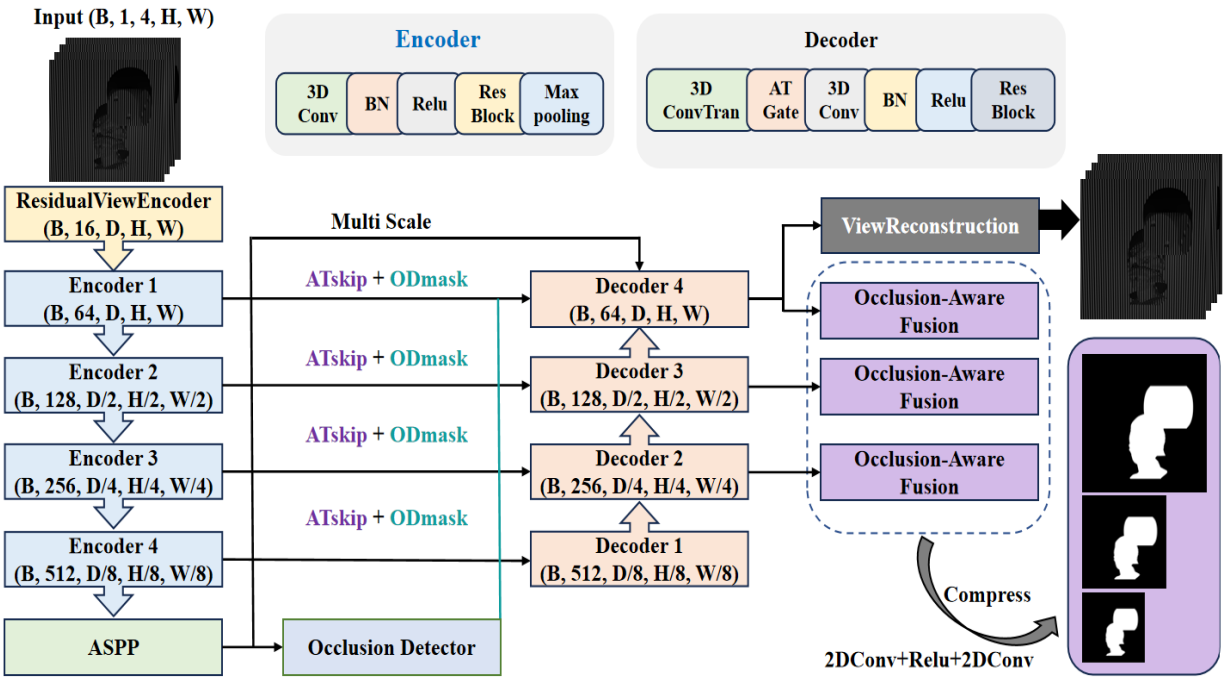

Fig. 1. Schematic overview of the ODMV-Net framework. The network utilizes a multi-task learning strategy with deep supervision to handle occlusions and reconstruct high-precision depth maps from multi-view inputs. Sub-modules are detailed in Fig. 2(a)-(d).

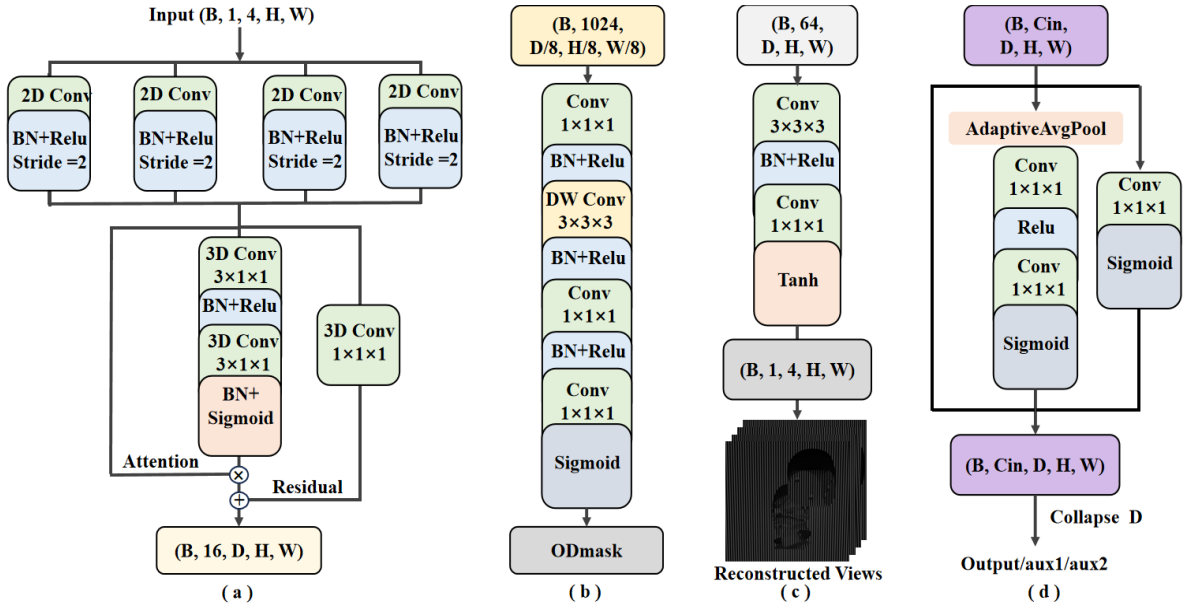

Fig. 2. Key components include: (a) ResidualViewEncoder, (b) Occlusion Detector, (c) View Reconstruction Branch, and (d) Occlusion-Aware Fusion.

A Self-Supervised View Reconstruction (SSVR) branch, illustrated in Fig. 2(c), is attached to the final decoding stage of the network. SSVR aims to reconstruct the original multi-view fringe inputs using a Tanh activation, thereby introducing an auxiliary self-supervised objective. By enforcing consistency between the reconstructed and input views, it acts as a structural regularizer, encouraging the shared latent representation to retain fine-grained geometric details that are essential for accurate depth estimation.

The OAF module in Fig. 2(d) is designed to convert multi-view 3D feature volumes into unified 2D depth representations across multiple output scales, including the Main, Aux1, and Aux2s. To achieve robust feature aggregation under occlusion, it employs a dual-branch attention mechanism that jointly models channel-view attention and spatial attention [25-26]. These complementary pathways adaptively reweight features based on their view-dependent reliability and spatial saliency. The attended features are then aggregated along the view dimension, producing coherent 2D feature maps that integrate complementary information from all viewpoints. The resulting representations are finally fed into scale-specific regression heads to generate high-fidelity depth maps.

To train the network, a composite multi-task loss function is adopted in conjunction with a deep supervision strategy. The overall objective is formulated as a weighted sum of depth reconstruction losses at multiple scales and an auxiliary self-supervised view reconstruction loss:

$$L_{total} = L_{main} + \lambda_1 L_{aux1} + \lambda_2 L_{aux2} + \lambda_3 L_{\text{view}} \quad (1)$$

where L_{main}, L_{aux1}, L_{aux2} are the depth losses from the Main, Aux1, and Aux2 scales, and L_{view} is the loss from the view reconstruction branch. The coefficients λ_1,λ_2 and λ_3 control the relative contributions of auxiliary supervision and structural regularization during training. In this work, the balancing weights were set to $\lambda_1 = 0.5$ and $\lambda_2 = 0.3$ to enforce hierarchical supervision, while $\lambda_3 = 0.2$ was selected to balance the gradient magnitudes between the depth estimation and view reconstruction tasks. These values were obtained based on empirical experiments.

The depth losses ($L_{main}, L_{aux1}, L_{aux2}$) are defined as a weighted combination of the Structural Similarity Index (SSIM) and the Mean Absolute Error (L_1):

$$L_{\text{depth}} = \alpha L_{\text{SSIM}} + (1-\alpha) L_1 \quad (2)$$

where α is set to 0.8 based on empirical experiments. The SSIM loss captures structural information and luminance consistency. The L_1 loss ensures pixel-wise accuracy by measuring the absolute difference between the predicted depth d and ground truth $\check{d}$. They are defined as:

$$L_{\text{SSIM}}(d, \check{d}) = 1 - \frac{(2\mu_d\mu_{\check{d}}+c_1)(2\sigma_{d\check{d}}+c_2)}{(\mu_d^2+\mu_{\check{d}}^2+c_1)(\sigma_d^2+\sigma_{\check{d}}^2+c_2)} \quad (3)$$

$$L_1(d, \check{d}) = \frac{1}{N}\sum_{i=1}^{N}|d_i - \check{d}_i| \quad (4)$$

where μ and σ denote the mean and variance of the depth maps, and c_1, c_2 are constants to maintain numerical stability.

The view reconstruction branch serves as an auxiliary task to enhance feature representation. It employs the Mean Squared Error (MSE) loss to minimize the intensity discrepancy between the reconstructed and input patterns:

$$L_{\text{view}}(I, \hat{I}) = \frac{1}{N}\sum_{i=1}^{N}(I_i - \check{I}_i)^2 \quad (5)$$

B. Hardware Configuration

As illustrated in Fig. 3(a-b), we modeled an FPP setup comprising a camera (4800×4800 pixels, 3.2μm×3.2μm pixel size, and 35.47mm×35.47mm FoV) surrounded by four projectors (1920×1080 pixels and 40mm×23mm FoV) arranged in a cross symmetric configuration. The projectors illuminate the target from four distinct directions: right, down, left, and up, corresponding to indices 1 to 4 in Fig. 3(d). Each projector is positioned at an angle of 26° relative to the central optical axis. The camera was calibrated with realistic lens distortion and synthetic noise to capture deformed fringe images under consistent illumination.

C. Dataset

To support robust, large-scale training and evaluation, we created a high-fidelity synthetic multi-view FPP dataset using Blender. This virtual data generation framework enables

accurate ground-truth collection and reproducible performance assessment [27-28]. The camera captures deformed fringe images rendered in 32-bit OpenEXR format to minimize quantization errors. Corresponding high-precision ground-truth depth maps were extracted directly from the Z-buffer. To ensure generalization, the dataset includes diverse topologies ranging from geometric primitives to complex industrial parts and freeform models from Thingi10K [29], as shown in Fig. 3(c). The final dataset comprises 37,296 images from 252 scenes. During training, only the highest-frequency fringe from each sequence was used. For efficient processing, all inputs were downsampled to 480×480 using trilinear interpolation. The data were partitioned into 80% for training, 10% for validation, and 10% for testing.

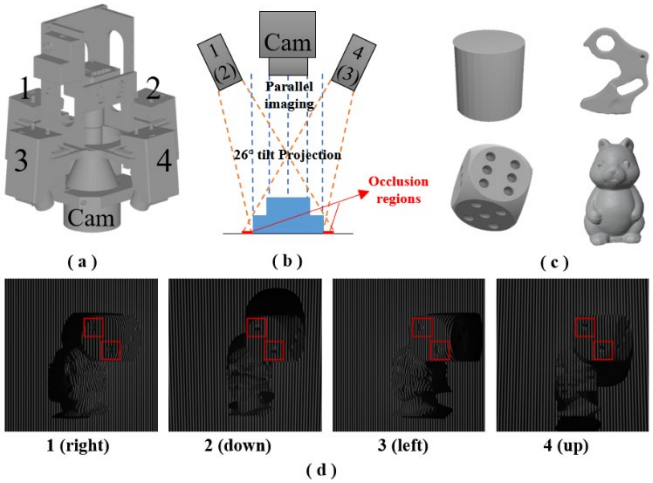

Fig. 3. Multi-view FPP simulation setup and dataset samples. (a-b) Hardware geometry and occlusion modeling (red horizontal lines); (c) Sample objects for training; (d) Orthogonal fringe inputs (1–4) showing projection-induced shadows (red boxes) addressed by the proposed network.

III. Experiment and Results

All experiments were conducted using a workstation equipped with an INTEL i7 CPU and an NVIDIA GeForce RTX 4090 GPU with 24GB VRAM, and all models were implemented in PyTorch 2.4.1 with CUDA 12.4. The network was trained for 200 epochs with a batch size of 2. Optimization was performed using the Adam optimizer, with the learning rate scheduled via cosine annealing with warm restarts to facilitate stable convergence. The ODMV-Net was systematically compared against representative state-of-the-art baselines, including UNet [30], hNet [31], UHRNet [32], and MTLNet [33]. To ensure a fair comparison, all competing methods were trained and evaluated under identical hardware, software, and training protocol.

A. View Experiments

The effectiveness of the proposed framework is clearly demonstrated by the four-view configuration (S=4). This balanced layout provides fully complementary observations across both horizontal and vertical directions, effectively resolving topological ambiguities caused by self-occlusion and shadowing. As a result, the four-view setting achieves the highest reconstruction fidelity, yielding the lowest errors (MAE = 0.0447 mm, MSE = 0.0403 mm, and RMSE = 0.1741 mm). Even with a reduced input of three views (S=3, omitting the 'down' view), the network continues to demonstrate its capability for effective multi-view information fusion. While the reconstruction precision marginally decreases relative to the S=4 setup, the model successfully leverages complementary features from the available perspectives to ensure a coherent and complete depth estimation. Further reducing the configuration to two views (S=2, Right and Up) leads to a more pronounced degradation in reconstruction quality. The absence of symmetric viewpoints results in persistent blind spots that cannot be adequately compensated through fusion. Finally, the single-view baseline (S=1, Right view only) exhibits the most severe performance degradation, where the lack of multi-angle observations causes substantial information loss and structural voids in self-occluded regions. These observations are quantitatively corroborated in Fig. 5, which presents depth profiles along the critical scanline indicated in Fig. 4.

B. Module Ablation Experiments

To rigorously quantify the contribution of each proposed component, we conducted a systematic ablation study by progressively augmenting a standard 3D U-Net backbone with additional modules. The results of this analysis are summarized in TABLE I.

The study begins with the baseline (vanilla 3D U-Net), which yields an initial MAE of 0.0926 mm. Incorporating residual blocks and attention gates (+RES+AT) results in a pronounced performance improvement, reducing the MAE to 0.0463 mm. This substantial gain indicates that residual connections effectively alleviate gradient degradation in deep feature hierarchies, while attention mechanisms enhance discriminative capacity by suppressing irrelevant background responses and emphasizing structurally salient regions.

TABLE I. Ablation Study on Network Components.

No.	Model	MAE (mm)	MSE (mm)	RMSE (mm)
1	Baseline	0.0926	0.1337	0.3442
2	+RB&AG	0.0463	0.0440	0.1840
3	+ASPP	0.0451	0.0432	0.1778
4	+ RVE	0.0461	0.0427	0.1776
5	+OCD&OAF	0.0452	0.0411	0.1747
6	+Multi-task	**0.0447**	**0.0403**	**0.1741**

Note: RB&AG: Residual Blocks [21-22] and Attention Gates [25-26]. RVE: Residual View Encoder. ASPP replaces intermediate convolutions [23-24]. OCD&OAF: Occlusion Detector and Occlusion-Aware Fusion.

The further integration of the ASPP module leads to an additional improvement, achieving an MAE of 0.0451 mm. This result highlights the importance of multi-scale contextual aggregation for resolving geometric ambiguities. The introduction of the RVE and OCD+OAF modules further stabilizes high-dimensional feature interactions and improves cross-view consistency. The residual connection within RVE enables explicit learning of multi-view alignment and integration, thereby producing more discriminative and view-consistent feature representations. Finally, extending the architecture to the full multi-task learning framework yields the best overall performance, with the MAE reaching 0.0447 mm.

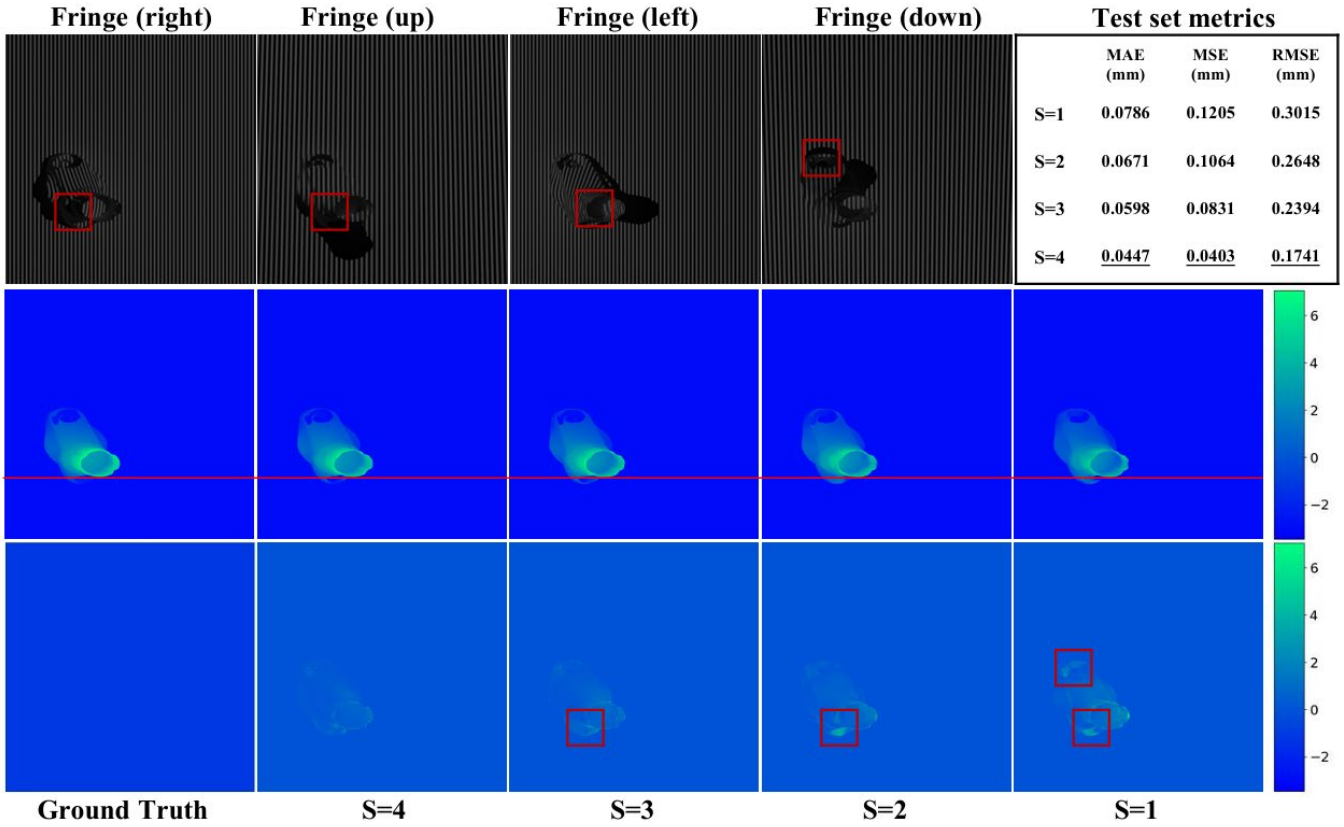

Fig. 4. Comparison of depth reconstruction accuracy. Top-to-bottom: Input fringes and test set metrics; GT versus predictions; and localized error maps. Red rectangles emphasize reconstruction artifacts in unilluminated regions due to complex object topology and occlusions.

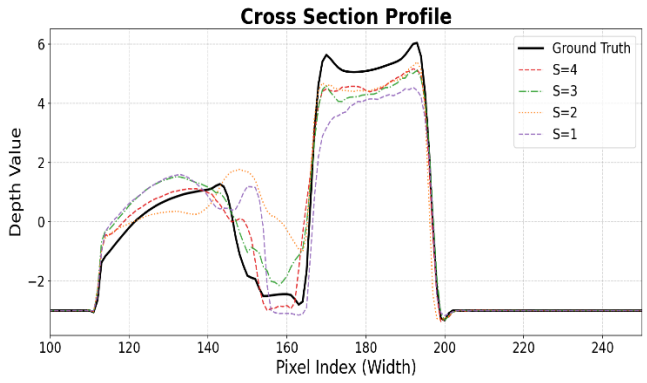

Fig. 5. Cross section depth profile of the red line on Fig. 4.

C. *Comparative Experiments*

We performed an extensive comparative analysis of the reconstructed depth maps and their corresponding absolute error distributions. As illustrated in Fig. 6 and summarized quantitatively in TABLE II, all five evaluated models are capable of recovering the global coarse morphology of the target object. However, pronounced differences emerge when assessing localized surface fidelity, boundary sharpness, and robustness in occluded regions.

The baseline hNet [31] exhibits significant degradation in fine structural details, producing fragmented and noisy surface reconstructions. This deficiency is further corroborated by its absolute error map, which contains widespread artifacts and elevated error magnitudes. UNet [30] achieves a modest improvement over hNet, yet its reconstructions remain affected by surface irregularities and inconsistent errors, particularly along sharp edges and structural boundaries.

In contrast, ODMV-Net consistently delivers superior reconstruction quality by exploiting a four-view fusion strategy that explicitly addresses occlusion and visibility constraints. While UHRNet [32] and MTLNet [33] produce visually smoother surfaces, their reliance on single-view input fundamentally limits their ability to resolve complex topological voids. This limitation is clearly visible in the regions highlighted by the red bounding boxes in Fig. 6, where both methods fail to recover depth in deeply occluded areas.

ODMV-Net effectively bridges these missing regions by integrating complementary observations across viewpoints, yielding more complete and geometrically consistent reconstructions. This advantage is further substantiated by the quantitative evaluation of a randomly selected depth profile across the object surface. As shown in Fig. 7, the depth variations along the arbitrary scanline reconstructed by ODMV-Net exhibit close agreement with the ground truth, whereas single-view methods display pronounced deviations and discontinuities due to unresolvable signal loss. These results collectively demonstrate the critical role of multi-view integration in achieving robust and accurate 3D reconstruction in challenging FPP scenarios.

Quantitatively, ODMV-Net establishes a new state-of-the-art, achieving an MAE of 0.0447 mm. This corresponds to substantial accuracy improvements of 58.73%, 69.15%, 57.39%, and 51.83% over UNet, hNet, UHRNet, and MTLNet, respectively. In addition, ODMV-Net reduces the MSE and RMSE to just 28.10% and 52.82%, respectively, of those obtained by the strongest competing model. These consistent gains across multiple error metrics highlight the proposed framework's superior precision, robustness, and reliability in challenging multi-view FPP reconstruction scenarios.

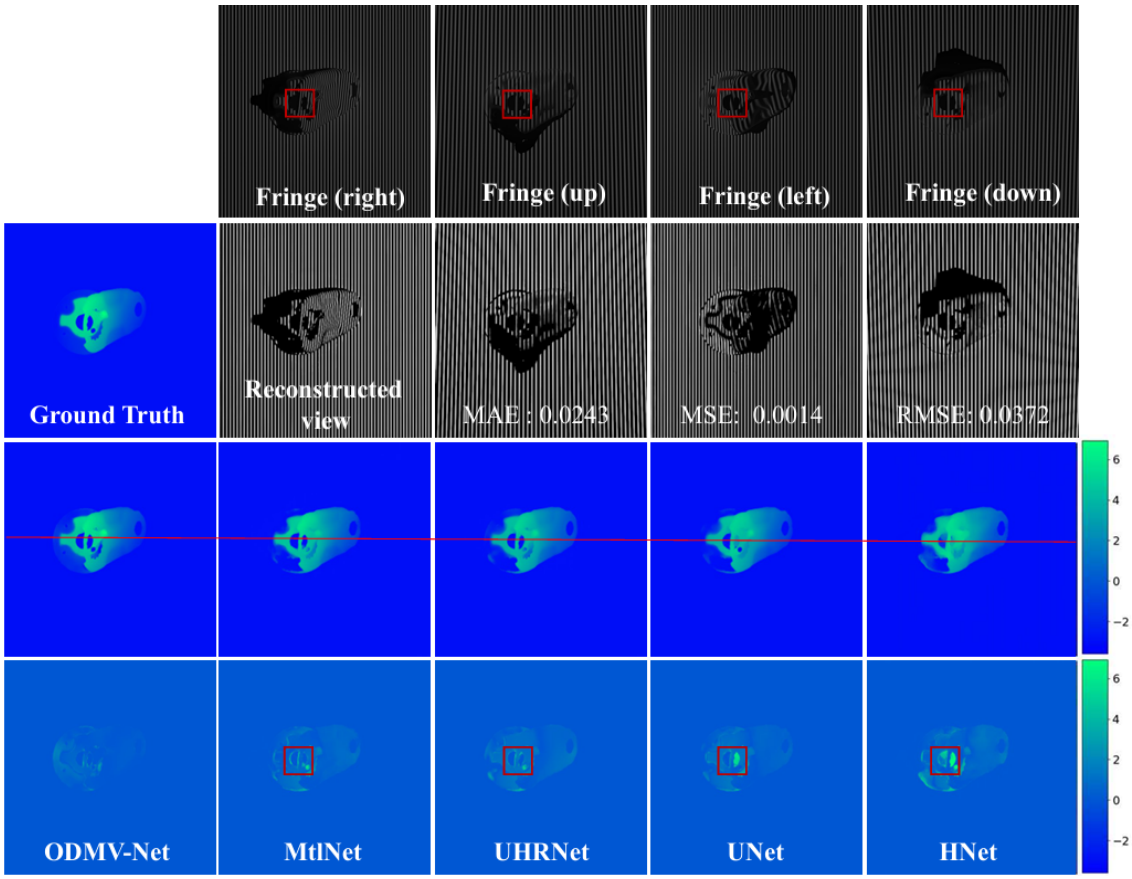

Fig. 6. Comparison of depth reconstruction results by different deep learning models. The figure displays input fringe patterns, ground truth, reconstructed views (with View Reconstruction branch error indicators of the test set), reconstructed depth maps, and the corresponding absolute error maps.

TABLE II. COMPARISON OF DIFFERENT MODELS.

Model	MAE (mm)	MSE (mm)	RMSE (mm)
hNet [31]	0.1449	0.2794	0.4792
UNet [30]	0.1083	0.2132	0.3979
UHRNet [32]	0.1049	0.1523	0.3316
MTLNet [33]	0.0928	0.1434	0.3296
ODMV-Net (Ours)	**0.0447**	**0.0403**	**0.1741**

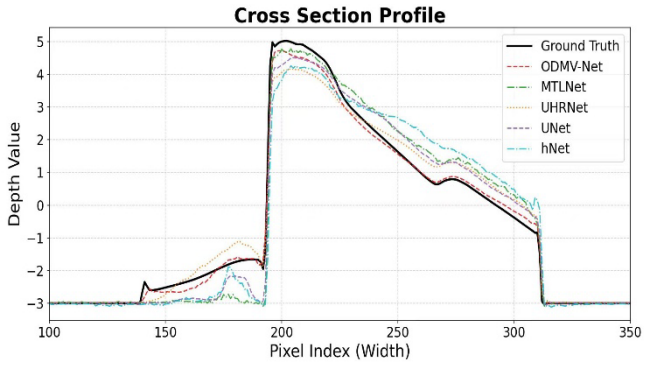

Fig. 7. Cross section depth profile comparison. The depth values are extracted along the red scanline marked in Fig. 6.

IV. CONCLUSION AND DISCUSSION

This study presents the development and validation of ODMV-Net, a framework designed to address the challenges of occlusion and topological incompleteness in FPP 3D reconstruction. By tightly integrating a multi-view fusion strategy with deep supervision and occlusion-aware learning, the proposed method exhibits strong robustness in recovering complex geometric boundaries and severely self-occluded regions. Extensive experiments demonstrate that a symmetric four-view configuration is critical for resolving geometric ambiguities, enabling ODMV-Net to consistently outperform conventional single-view methods and achieve a substantially reduced mean absolute error of 0.0447 mm. Collectively, these findings establish ODMV-Net as a reliable and extensible multi-view architecture for high-precision surface reconstruction in advanced fringe projection profilometry applications.

Despite the demonstrated effectiveness of the ODMV-Net, several limitations remain. First, the high-dimensional nature of volumetric 3D feature processing, combined with multi-view fusion, incurs a substantial computational overhead and elevated GPU memory requirements. The model has a large number of parameters (112,560,408), which complicates deployment. This computational burden may limit real-time deployment on resource-constrained platforms such as embedded or edge devices.

Second, although large-scale synthetic datasets enable controlled training and benchmarking, the generalization of the network to real-world scenarios with imperfect calibration, sensor noise, and domain discrepancies remains an open challenge. In particular, the robustness of the framework under extreme optical conditions—such as highly reflective or transparent surfaces, low-contrast fringe patterns, and dynamic scenes—has not yet been fully explored.

Third, the present implementation assumes a fixed and symmetric four-view acquisition layout. While it is effective in resolving occlusions and geometric ambiguities, it inherently limits flexibility in unstructured or space-constrained environments, where sensor placement may need to be adapted dynamically due to physical constraints or task-specific requirements.

Future work will focus on addressing these challenges through several complementary directions. To mitigate computational and memory constraints, we plan to investigate model compression and acceleration strategies, including knowledge distillation, structured pruning, and lightweight attention mechanisms, with the goal of enabling low-latency inference suitable for in-line inspection. To reduce dependence on dense ground-truth annotations, we will explore self-supervised and weakly supervised learning paradigms that exploit cross-view photometric and geometric consistency as intrinsic supervisory signals. Finally, extending the current 3D convolutional architecture to incorporate temporal modeling—through recurrent units, temporal convolutions, or spatiotemporal attention—will allow the system to achieve high-fidelity reconstruction of moving or deforming objects in dynamic measurements.

ACKNOWLEDGMENT

The study was funded by Guangdong Basic and Applied Basic Research Major Project (2023B0303000009), Shenzhen University (2023YQ015) and China National Nuclear Corporation (CNNCLCKY-2024-072).

REFERENCES

[1] S. Feng, L. Zhang, C. Zuo, T. Tao, Q. Chen, and G. Gu, "High dynamic range 3D measurements with fringe projection profilometry: a review," *Meas. Sci. Technol.*, vol. 29, no. 12, p. 122001, 2018.

[2] C. Zuo, S. Feng, L. Huang, T. Tao, W. Yin, and Q. Chen, "Phase shifting algorithms for fringe projection profilometry: a review," *Opt. Lasers Eng.*, vol. 109, pp. 23–59, 2018.

[3] J. Yu, N. Gao, Z. Zhang, and Z. Meng, "High sensitivity fringe projection profilometry combining optimal fringe frequency and optimal fringe direction," *Opt. Lasers Eng.*, vol. 129, p. 106068, 2020.

[4] S. Lv, X. Zhang, G. Zhao, Z. Zhang, and J. Zhang, "Fringe projection profilometry method with high efficiency, precision, and convenience: theoretical analysis and development," *Opt. Express*, vol. 30, no. 19, pp. 33515–33537, 2022.

[5] Z. H. Zhang, "Review of single-shot 3D shape measurement by phase calculation-based fringe projection techniques," *Opt. Lasers Eng.*, vol. 50, no. 8, pp. 1097–1106, 2012.

[6] Z. Zhu, M. Li, F. Zhou, and D. You, "Stable 3D measurement method for high dynamic range surfaces based on fringe projection profilometry," *Opt. Lasers Eng.*, vol. 166, p. 107542, 2023.

[7] S. Feng, Q. Chen, G. Gu, T. Tao, L. Zhang, Y. Hu, et al., "Fringe pattern analysis using deep learning," *Adv. Photon.*, vol. 1, no. 2, p. 025001, 2019.

[8] W. Yin, Y. Che, X. Li, M. Li, Y. Hu, S. Feng, et al., "Physics-informed deep learning for fringe pattern analysis," *Opto-Electron. Adv.*, vol. 7, no. 1, p. 230034-1, 2024.

[9] Y. Rivenson, Y. Zhang, H. Günaydın, D. Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," *Light Sci. Appl.*, vol. 7, no. 2, p. 17141, 2018.

[10] K. Yan, Y. Yu, C. Huang, L. Sui, K. Qian, and A. Asundi, "Fringe pattern denoising based on deep learning," *Opt. Commun.*, vol. 437, pp. 148–152, 2019.

[11] G. E. Spoorthi, S. Gorthi, and R. K. S. S. Gorthi, "PhaseNet: a deep convolutional neural network for two-dimensional phase unwrapping," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 54–58, 2019.

[12] W. Yin, Q. Chen, S. Feng, T. Tao, L. Huang, M. Trusiak, et al., "Temporal phase unwrapping using deep learning," *Sci. Rep.*, vol. 9, no. 1, p. 20175, 2019.

[13] R. C. Machineni, G. E. Spoorthi, K. S. Vengala, S. Gorthi, and R. K. S. S. Gorthi, "End-to-end deep learning-based fringe projection framework for 3D profiling of objects," *Comput. Vis. Image Underst.*, vol. 199, p. 103023, 2020.

[14] X. Zhu, T. Lan, Y. Zhao, H. Wang, and L. Song, "End-to-end color fringe depth estimation based on a three-branch U-net network," *Appl. Opt.*, vol. 63, no. 28, pp. 7465–7474, 2024.

[15] Z. Zhang, Y. Li, Z. Han, C. Zhang, X. Liang, and X. Li, "An end-to-end network for 3D reconstruction from multi-frequency fringe patterns," in *Proc. 26th China Conf. Syst. Simul. Technol. Appl. (CCSSTA)*, 2025, pp. 474–479.

[16] C. Wang, P. Zhou, and J. Zhu, "Deep learning-based end-to-end 3D depth recovery from a single-frame fringe pattern with the MSUNet++ network," *Opt. Express*, vol. 31, no. 20, pp. 33287–33298, 2023.

[17] G. Zhang, Y. Liu, Q. Yao, H. Deng, H. Zhao, Z. Zhang, and S. Yang, "Multi-view fringe projection profilometry for surfaces with intricate structures and high dynamic range," *Opt. Express*, vol. 32, no. 11, pp. 19146–19162, 2024.

[18] A. Dickins, T. Widjanarko, D. Sims-Waterhouse, A. Thompson, S. Lawes, N. Senin, and R. Leach, "Multi-view fringe projection system for surface topography measurement during metal powder bed fusion," *J. Opt. Soc. Am. A*, vol. 37, no. 9, pp. B93–B105, 2020.

[19] Y. Ren, W. Tao, and H. Zhao, "Multi-view fringe projection profilometry based on phase texture and U-Net," *Opt. Express*, vol. 32, no. 16, pp. 27690–27709, 2024.

[20] M. Wang, Y. Yin, D. Deng, X. Meng, X. Liu, and X. Peng, "Improved performance of multi-view fringe projection 3D microscopy," *Opt. Express*, vol. 25, no. 16, pp. 19408–19421, 2017.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.

[22] C. Wu, Z. Qiao, N. Zhang, X. Li, J. Fan, H. Song, et al., "Phase unwrapping based on a residual en-decoder network for phase images in Fourier domain Doppler optical coherence tomography," *Biomed. Opt. Express*, vol. 11, no. 4, pp. 1760–1771, 2020.

[23] Y. Qiu, Y. Liu, Y. Chen, J. Zhang, J. Zhu, and J. Xu, "A2SPPNet: attentive atrous spatial pyramid pooling network for salient object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 1991–2006, 2022.

[24] N. A. Mohamed, M. A. Zulkifley, and S. R. Abdani, "Spatial pyramid pooling with atrous convolutional for mobilenet," in *Proc. IEEE Student Conf. Res. Dev. (SCOReD)*, pp. 333–336, 2020.

[25] W. Deng, Q. Shi, and J. Li, "Attention-gate-based encoder–decoder network for automatical building extraction," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2611–2620, 2021.

[26] J. Nodirov, A. B. Abdusalomov, and T. K. Whangbo, "Attention 3D U-Net with multiple skip connections for segmentation of brain tumor images," *Sensors*, vol. 22, no. 17, p. 6501, 2022.

[27] H. Nguyen, E. Novak, and Z. Wang, "Accurate 3D reconstruction via fringe-to-phase network," *Measurement*, vol. 190, p. 110663, 2022.

[28] X. Zhu, Z. Zhang, L. Hou, L. Song, and H. Wang, "Light field structured light projection data generation with Blender," in *Proc. 3rd Int. Conf. Comput. Vis. Image Deep Learn. Int. Conf. Comput. Eng. Appl. (CVIDL & ICCEA)*, pp. 1249–1253, 2022.

[29] Q. Zhou and A. Jacobson, "Thingi10k: a dataset of 10,000 3d-printing models," arXiv preprint arXiv:1605.04797, 2016.

[30] H. Nguyen, Y. Wang, and Z. Wang, "Single-shot 3D shape reconstruction using structured light and deep convolutional neural networks," *Sensors*, vol. 20, no. 13, p. 3718, 2020.

[31] H. Nguyen, K. L. Ly, T. Tran, Y. Wang, and Z. Wang, "hNet: single-shot 3D shape reconstruction using structured light and h-shaped global guidance network," *Results Opt.*, vol. 4, p. 100104, 2021.

[32] Y. Wang, C. Zhou, X. Qi, and H. Li, "UHRNet: a deep learning-based method for accurate 3D reconstruction from a single fringe-pattern," *J. Mod. Opt.*, vol. 70, no. 11, pp. 707–722, 2023.

[33] X. Zhong, J. Huang, Y. Li, J. Chen, Y. Tian, and Z. Wu, "Single-shot fringe projection profilometry based on multi-task learning: efficient depth reconstruction without explicit system calibrations," *Phys. Scr.*, vol. 100, no. 2, p. 026005, 2025.