Advanced Devices & Instrumentation
A SCIENCE PARTNER JOURNAL

RESEARCH ARTICLE

Deep Learning Enhanced Dynamic 3-Dimensional Shape Measurement Using Single-Shot Spatial Multiplexing and Transformer-Based Phase Retrieval

Yixuan Li[1,2,3†], Yile Xiao[1,2,3†], Jiaming Qian[1,2,3*], Shijie Feng[1,2,3*], Qian Chen[1,2,3*], and Chao Zuo[1,2,3*]

[1]Smart Computational Imaging Laboratory (SCILab), School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China. [2]Smart Computational Imaging Research Institute (SCIRI) of Nanjing University of Science and Technology, Nanjing, Jiangsu 210019, China. [3]Jiangsu Key Laboratory of Spectral Imaging and Intelligent Sense, Nanjing 210094, China.

*Address correspondence to: jiaming_qian@njust.edu.cn (J.Q.); shijiefeng@njust.edu.cn (S.F.); chenqian@njust.edu.cn (Q.C.); zuochao@njust.edu.cn (C.Z.)
†These authors contributed equally to this work.

Citation: Li Y, Xiao Y, Qian J, Feng S, Chen Q, Zuo C. Deep Learning Enhanced Dynamic 3-Dimensional Shape Measurement Using Single-Shot Spatial Multiplexing and Transformer-Based Phase Retrieval. *Adv. Devices Instrum.* 2025;6:Article 0095. https://doi.org/10.34133/adi.0095

Submitted 19 November 2024
Revised 7 February 2025
Accepted 25 February 2025
Published 4 November 2025

Real-time, high-precision 3-dimensional (3D) imaging is essential for applications such as industrial inspection, robotic navigation, and human–computer interaction. Fringe projection profilometry (FPP), a widely used structured light method, achieves high spatiotemporal resolution by rapidly projecting and processing of fringe patterns. However, traditional multi-frame FPP methods are hindered by motion-induced artifacts and computational bottlenecks, limiting their applicability in dynamic environments. In this work, we propose a multiplexed structured light 3D measurement method that integrates a physics-based Transformer framework to minimize the number of projection patterns required for precise single-snapshot measurements. This method extracts accurate phase information from a single fringe image, enabling artifact-free, high-resolution 3D surface reconstruction. By combining low-frequency triangular waves with high-frequency sinusoidal fringes, we ensure unambiguous phase retrieval, providing the deep neural network with reliable inputs. The Transformer-based network leverages superior global information capture and multi-scale feature learning capabilities for robust fringe analysis and phase unwrapping, thereby enhancing the accuracy and generalization of depth prediction. Experimental evaluations demonstrate that our method outperforms traditional single-frame phase retrieval techniques and other deep learning-based methods in terms of precision and robustness. Dynamic measurements of complex objects with various materials further validate its potential for high-speed, real-time 3D imaging in intelligent manufacturing and augmented reality.

Introduction

High-speed 3-dimensional (3D) imaging is essential for applications demanding real-time data acquisition and processing, such as industrial manufacturing, medical diagnosis, cultural heritage preservation, robotic navigation, and motion analysis [1–5]. In these domains, rapidly and accurately capturing the dynamic 3D surface information is essential for tasks like quality control, object recognition, and interaction with the environment. As industry moves toward automation and real-time monitoring, the need for high-efficiency, precise 3D measurement techniques has increased accordingly. Structured light fringe projection profilometry (FPP) is a commonly employed optical technique for 3D surface reconstruction, valued for its high accuracy and simplicity in implementation [6–9]. Traditional FPP techniques involve projecting multiple digitally encoded structured light patterns (typically sinusoidal fringes) onto the sample surface and recording the deformed images from camera perspectives. The phase information encoded in these patterns is extracted and unwrapped to reconstruct the object's surface geometry. While multi-frame FPP achieves high precision in static scenes, it faces severe challenges in dynamic scenarios, where multiple-frame acquisition introduces motion-induced artifacts and phase mismatches, leading to substantial errors that hinder its applicability to real-time or high-speed measurements.

Single-frame 3D imaging mitigates ambient lighting fluctuations across frames and overcomes motion-induced interference [10–12]. Capturing all necessary information in one shot enables accurate measurements of moving or vibrating targets, which is critical in settings such as automated industrial inspection. Consequently, achieving high-precision 3D reconstruction from a single frame has been a longstanding objective in structured light imaging. Traditional single-shot approaches like Fourier transform profilometry (FTP) [13,14] compute

phases from a deformed fringe map via Fourier transformations and band-pass filtering in the frequency domain. However, FTP often assumes relatively smooth surfaces and tolerates only moderate height variations; surfaces featuring large discontinuities or intricate geometries introduce strong high-frequency noise that can degrade reconstruction accuracy. Advances in artificial intelligence (AI) have extended deep learning (DL) frameworks to computational imaging [15,16], including optical 3D measurements [17,18]. These developments open promising avenues for single-frame, high-precision 3D imaging in complex dynamic environments. Deep neural networks can extract precise wrapped phase data from a single-pattern input [19–21] and, in some cases, predict the absolute phase or depth [22,23]. However, challenges remain for single-frame methods when facing complex surface features (e.g., contours or textures), varying illumination, and noise, as deriving an accurate mapping from the intensity image to the phase or depth can be difficult in purely data-driven networks.

Spatial multiplexing has shown promise in producing high-precision, single-shot 3D measurements by embedding additional coding information into a single projected pattern [24–27]. Strategies involving frequency multiplexing, color encoding, and composite fringe patterns allow the system to recover absolute phase from a single capture [28–30]. Although spatial multiplexing can encode multiple fringe patterns into one image–thereby maximizing use of spatial dimensions and boosting noise resistance—measurement accuracy sometimes suffers from increased pattern complexity, surface reflectivity variations, or the intricate decoding algorithms required. Recent work has explored convolutional neural networks (CNNs) [31–33] to improve the accuracy and robustness of phase retrieval. However, multi-scale or single-frame settings often demand global context to resolve phase ambiguities, and CNNs are typically constrained by their local receptive fields. Transformers, originally developed for natural language processing, employ self-attention mechanism capable of capturing both local and long-range dependencies [34–36]. Applied to image-based tasks, Transformers can learn nuanced global structures and fine-scale details, showing promising performance in applications such as classification, segmentation, and depth estimation.

In this work, a novel single-shot spatial multiplexed structured light 3D imaging approach is proposed, integrating a physics-based Transformer framework. Our encoding strategy embeds low-frequency triangular-wave information within high-frequency sinusoidal fringes, creating a composite pattern that encodes unambiguous phase data. The triangular waves serve as robust spatial markers, assisting in resolving phase ambiguities inherent in high-frequency fringes and furnishing the neural network with reliable cues for learning. A Transformer-based architecture with self-attention mechanism then leverages these cues to capture both local and global features, enhancing fringe analysis and phase unwrapping. This design shows superior performance relative to traditional CNN-based architectures such as UNet, particularly for single-frame phase retrieval tasks involving complex or dynamic scenes. Experimental evaluations confirm that this triangular-wave-embedded strategy outperforms standard single-sinusoidal and dual-frequency fringe techniques. Furthermore, the proposed Transformer network consistently achieves higher accuracy and robustness in retrieving phase information, thereby delivering artifact-free 3D reconstructions of moving and intricately shaped objects. Measurement accuracies reach approximately 65 μm, demonstrating that this approach meets the stringent precision needs of industrial or high-precision applications.

Materials and Methods

Design of the spatial multiplexing coding strategy

In FPP, a single fringe image is employed to eliminate motion-induced interference during phase retrieval. When a standard sinusoidal fringe pattern is projected, the intensity captured at position (x,y) often takes the form

$$I_{fri}(x,y) = A^c(x,y) + B^c(x,y)\cos\left[\phi(x,y)\right], \tag{1}$$

where A^c represents the background illumination, B^c indicates the modulation amplitude, and ϕ corresponds to the phase associated with the surface height of the object. Extracting accurate phase information from a single sinusoidal fringe image remains challenging in complex scenes due to limited intensity information and inherent fringe ambiguity [6,29], often leading to erroneous pixel correspondences and phase unwrapping failures. To address these limitations, a spatial multiplexing coding strategy is employed, embedding multiple fringe patterns into one composite image. This additional information aids demodulation and ensures unambiguous phase unwrapping. Figure 1B demonstrates a composite encoding approach that superimposes 2 fringe patterns of distinct frequencies. In this dual-frequency design, a relatively low-frequency fringe provides auxiliary information to assist the high-frequency fringe in determining the absolute phase and the correct fringe order. However, this approach can be susceptible to spectrum aliasing: The low-frequency component (f_L) may interfere with the high-frequency spectrum (f_H), leading to ambiguity. To overcome such constraints, we propose an enhanced composite encoding strategy by embedding a triangular wave into the sinusoidal fringes, as depicted in Fig. 1C. The projected composite fringe pattern is formulated as

$$I^p_{tri-fri}(x,y) = A^p + B^p\cos\left(2\pi f_0 x\right) + \alpha\cdot T(x,y), \tag{2}$$

where A^p denotes the average intensity of the projector, B^p represents the amplitude of the sinusoidal fringe pattern with spatial frequency f_0, $T(x,y)$ is the triangular wave function with wavelength λ_{tri}, and coefficient α adjusts the contribution of the triangular wave to the composite pattern. The triangular wave function

$$T(x,y) = \frac{2}{\lambda_{tri}}\left|\mathrm{mod}\left(x+\frac{\lambda_{tri}}{2},\lambda_{tri}\right)-\frac{\lambda_{tri}}{2}\right| \tag{3}$$

has wavelength λ_{tri}, with $\mathrm{mod}(\cdot)$ indicating the modulus operation. The periodic peaks of the triangular wave serve as spatial markers that assist phase unwrapping without substantially degrading the high-frequency phase information.

Figure 1B and C shows the corresponding spectral distributions. In Fig. 1B, the low-frequency component (f_L) has a higher amplitude than the high-frequency term (f_H), whereas in Fig. 1C, the amplitude of f_L is smaller than that of f_H. This amplitude relationship indicates that the embedded triangular wave has a minimal impact on the high-frequency sinusoidal fringe, thereby enhancing the accuracy of high-frequency

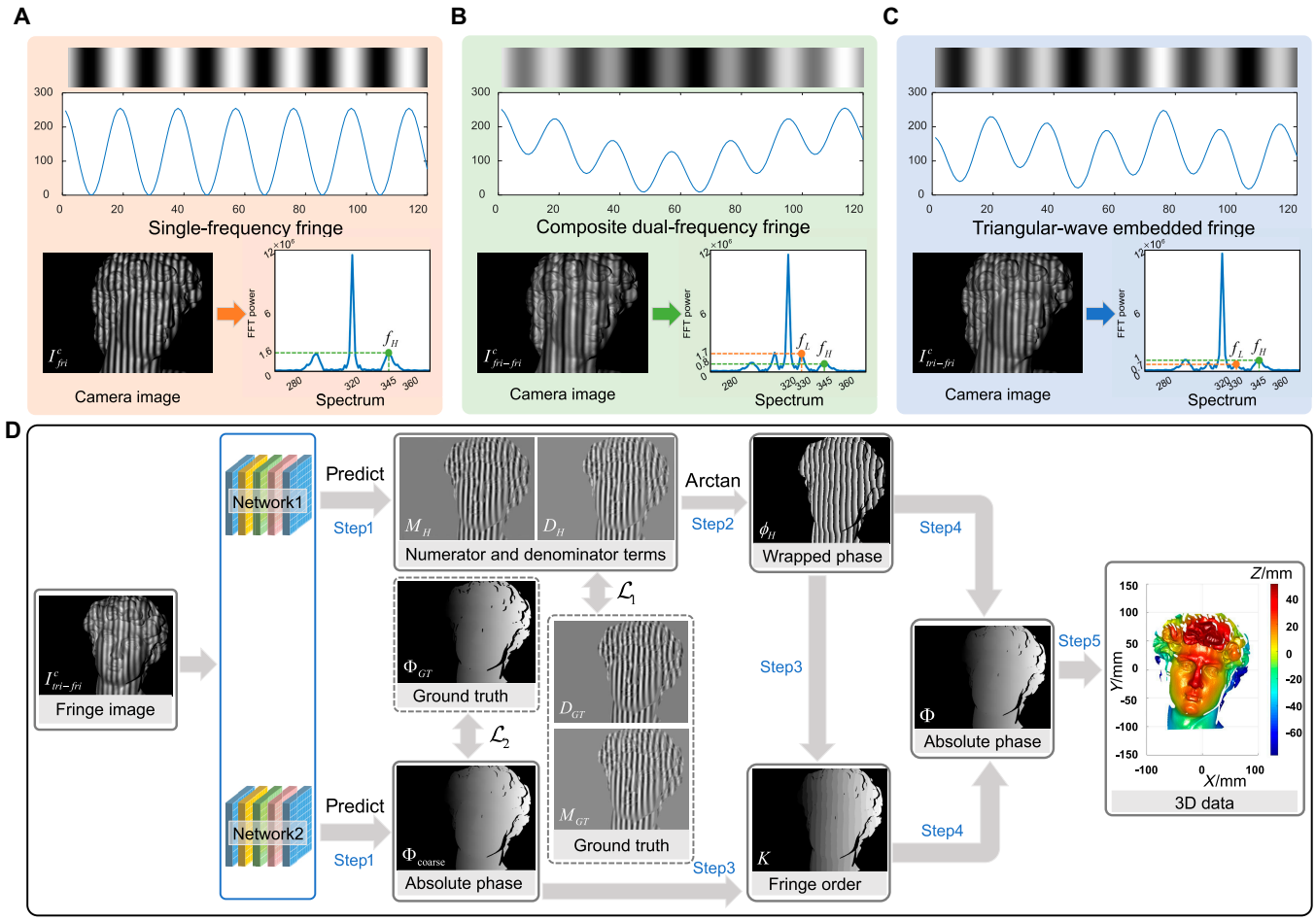

Fig. 1. Single-frame composite structured light 3D imaging utilizing a Transformer network with self-attention mechanism. (A) Standard single high-frequency fringe, (B) composite dual-frequency fringe [29], and (C) triangular-wave-embedded fringe, each accompanied by their respective projection pattern sequences, camera images, and corresponding spectral cross-sectional intensity distributions. In (B) and (C), f_L and f_H denote the low- and high-frequency components, respectively, which are carefully selected to prevent spectrum aliasing. (D) Flowchart of the proposed phase retrieval and 3D reconstruction process using the enhanced Swin-Unet Transformer-based network. Step 1: The captured single-frame triangular-wave-embedded fringe image ($I^c_{tri-fri}$) is simultaneously processed by 2 networks: Network 1 predicts the numerator (M_H) and denominator (D_H) terms of the wrapped phase, while Network 2 estimates the coarse absolute phase (Φ_{coarse}). Step 2: The numerator and denominator terms predicted by Network 1 are then used to compute the wrapped phase ϕ_H via the arctangent function as defined in Eq. (5). Step 3: The output from Network 2, combined with the calculated wrapped phase, predicts fringe order K. Step 4: Using Eq. (6), the recovered absolute phase Φ is calculated from ϕ_H and K. Step 5: The 3D data are generated from this phase, showing the reconstructed surface of the object. Two loss functions, $\mathcal{L}_1$ and $\mathcal{L}_2$, are employed to optimize the predictions from Network 1 and Network 2, respectively, by comparing them with the ground truth. The overall process enables precise reconstruction of 3D point cloud data from a single camera image.

phase retrieval. By incorporating spatial positioning information from the low-frequency triangular wave into the high-frequency sinusoidal fringes, the composite pattern effectively carries both high-resolution phase details and unambiguous spatial encoding. This encoding significantly improves the DL network's ability to retrieve precise and unambiguous phase information from a single captured image. We further validate this theoretical analysis in the experimental results. To facilitate reliable phase unwrapping, it is necessary to select the wavelengths of the primary sinusoidal fringe (λ_0) and the auxiliary triangular wave (λ_{tri}) such that

$$\mathrm{LCM}\left(\lambda_0, \lambda_{tri}\right) \geq W, \tag{4}$$

where LCM(·) represents the least common multiple (LCM), and W represents the lateral resolution of the projector. Ensuring that the LCM of λ_0 and λ_{tri} exceeds W allows the wrapped phase to be unambiguously resolved across all pixels in the image, thus providing robust phase unwrapping under real-world operating conditions.

Single-shot triangular-wave-embedded FPP with Transformers

Accurate phase retrieval is fundamental to achieving high-precision 3D reconstruction in FPP. Figure 1D illustrates the workflow of our single-frame triangular-wave-embedded fringe projection and 3D reconstruction method, which employs a Transformer-based network. The core design leverages DL to extract 3 key components from the single-shot multiplexed image: the sine term (numerator) M_H, the cosine term (denominator) D_H, and a coarse absolute phase Φ_{coarse}. Here, M_H and D_H correspond to the numerator and denominator of the tangent function used to compute the wrapped phase ϕ_H, as defined by

$$\phi_H = \tan^{-1}\left(\frac{M_H}{D_H}\right) = \tan^{-1}\left[\frac{\sum_{n=1}^{12} I_{Hn}(x,y)\sin(2\pi(n-1)/N)}{\sum_{n=1}^{12} I_{Hn}(x,y)\cos(2\pi(n-1)/N)}\right], \quad (5)$$

where I_{Hn} $(n = 1, 2, \ldots, 12)$ denotes the 12-step phase-shifted images captured under high-wavelength fringes. Within our framework, Network 1 predicts M_H and D_H directly from the triangular-wave-embedded fringe image, thereby avoiding the abrupt phase jumps often observed in wrapped phase outputs. Additionally, by incorporating the physical model of FPP phase analysis, this approach counteracts interference due to surface reflectivity variations, allowing robust phase retrieval and unambiguous unwrapping even in complex environments. This reliability serves as a solid foundation for subsequent 3D reconstruction.

Once M_H and D_H have been obtained from Network 1, the wrapped phase ϕ_H follows from Eq. (5). Although this step achieves high-accuracy phase determination, the 2π wrapping effect still renders the phase ambiguous. To resolve this ambiguity, Network 2 predicts a coarse estimate of the absolute phase Φ_{coarse}, offering an initial approximation of the phase distribution across the entire scene. The final absolute phase Φ is then computed by combining the wrapped phase ϕ_H from Network 1 with the coarse phase from Network 2, according to

$$\Phi = \phi_H + 2\pi K = \phi_H + 2\pi \cdot \text{Round}\left[\left(\Phi_{\text{coarse}} - \phi_H\right)/2\pi\right]. \quad (6)$$

Here, $\text{Round}\left[\left(\Phi_{\text{coarse}} - \phi_H\right)/2\pi\right]$ identifies the correct fringe order K, thereby adjusting ϕ_H to yield the unwrapped absolute phase Φ. The coarse phase Φ_{coarse} thus distinguishes different periods of the wrapped phase, ensuring accurate phase unwrapping. This final unwrapped phase reliably represents the object's surface without phase ambiguities. After precisely retrieving and refining Φ, the 3D geometry of the sample can be reconstructed using phase-to-depth mapping functions alongside system calibration parameters.

Transformer-based neural network architecture

To improve the precision and robustness of phase retrieval and unwrapping, we employ a Transformer-based neural network (referred to as the enhanced Swin-Unet), which leverages advanced self-attention mechanisms to effectively capture local and global feature information of feature maps. Figure 2 shows the complete architecture of the enhanced Swin-Unet, while Fig. 2A illustrates its basic encoder–decoder structure. In the enhanced Swin-Unet, we replace conventional convolutional layers with Swin Transformers by converting the smallest processing unit from individual pixels into $P \times P$ patches via the Partition block (Fig. 2B). Each $P \times P$ patch encapsulates the local structure and contextual information more effectively than single pixels, enabling the self-attention mechanism to capture global dependencies among patches rather than being restricted by the purely local receptive fields of typical CNN. Moreover, patch-level processing mitigates sensitivity to random noise by averaging intensity fluctuations within each patch, thereby improving both the robustness and accuracy of feature extraction and subsequent phase retrieval. Consequently, this patch-level Transformer design not only reduces computational complexity but also preserves essential feature relationships. Figure 2C depicts the structure of the Swin Transformer block, consisting of a sequence of modular components via residual connections. The block begins with layer normalization in each submodule to stabilize and normalize feature distributions. It utilizes windowed multi-head self-attention (W-MSA) to capture information within local regions, balancing computational efficiency and contextual understanding. Subsequently, a multi-layer perceptron (MLP) with a Gaussian error linear unit (GELU) activation function further refines feature representations and introduces nonlinearity to enhance the expressiveness of the model. The block also employs shifted window multi-head self-attention (SW-MSA), allowing the network to effectively catch relationships across local regions while mitigating the limitations of window boundaries. Residual connections in this structure ensure the preservation of crucial input features and improve the stability and efficiency of gradient flow during training.

The proposed network architecture comprises an encoder, a bottleneck module, and a decoder. Within the encoder, 2 successive Swin Transformer blocks extract features while maintaining the original spatial resolution. A Patch Merging module (Fig. 2D) then halves the spatial resolution while doubling the number of channels, repeated 3 times to capture multi-scale features essential for accurate phase retrieval. The bottleneck module, comprising 2 Swin Transformer blocks, further reduces the feature map to its smallest spatial size and highest semantic level, facilitating a seamless transition between the encoder and decoder. In the decoder, the Patch Expand module (Fig. 2E) upsamples the feature map, doubling its resolution and halving the channel count in a manner complementary to Patch Merging. Skip connections then fuse the outputs of the Patch Expand module with the corresponding encoder feature maps, ensuring that both fine-grained details and high-level context are retained. Ultimately, the Patch Reverse module restores the final feature map to the original resolution, producing the numerator, denominator, and "coarse" absolute phase terms needed for unambiguous phase retrieval.

In addition, the self-attention mechanism is improved by replacing the fixed square window in the original network with a more adaptable and larger window, as illustrated in Fig. 2F. In standard configurations, the minimum feature map size of $\frac{H}{32} \times \frac{W}{32}$ restricts the window attention mechanism, often denoted by c. Here, c must meet the conditions $n \times A = \frac{H}{32}$ and $m \times A = \frac{W}{32}$, where n and m are the numbers of vertical and horizontal windows, respectively. In this work, the attention window size is aligned with dimensions after the final downsampling, set to $A = \frac{H}{32}$ and $B = \frac{W}{32}$. This modification significantly broadens the area each window covers, establishing stronger pixel correlations and capturing a wider scope of global contextual information for modeling long-range dependencies. Combining this Transformer-based enhanced Swin-Unet network with our spatial multiplexing encoding strategy, we can effectively decode precise, unambiguous phase information from complex scenes using only one-shot image. The combination of windowed self-attentions and multi-scale feature learning allows the network to utilize both global and local features in fringe analysis and phase unwrapping tasks, thus improving the accuracy of the outputs. The proposed Transformer-based network outperforms traditional CNN architectures (e.g., Unet)

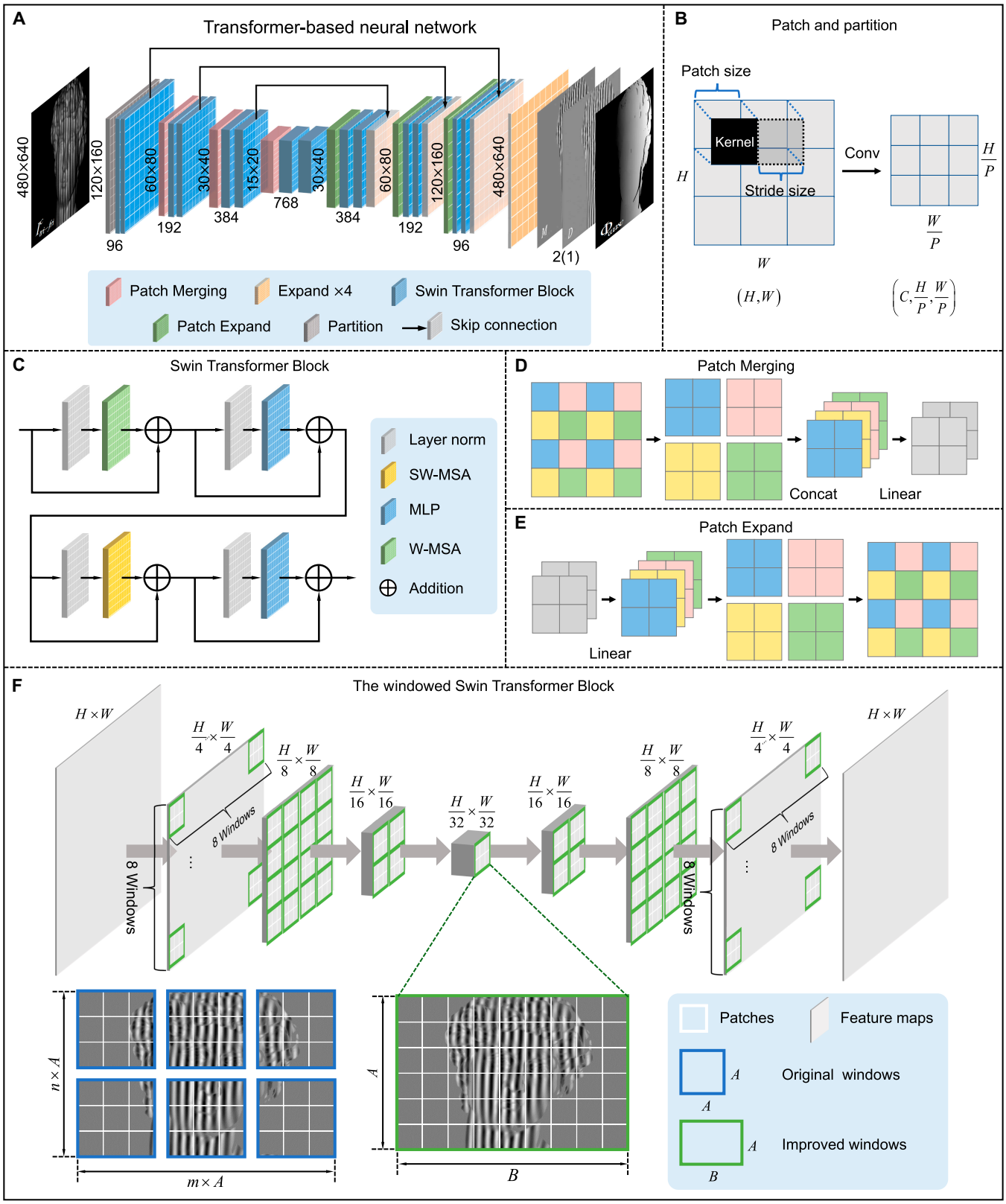

Fig. 2. The Transformer-based network architecture (the enhanced Swin-Unet). (A) Basic structure of the enhanced Swin-Unet. (B) Partition module. (C) Swin Transformer block. (D) Patch Merging module. (E) Patch Expand module. (F) The window of Swin Transformer Block.

in terms of prediction accuracy and robustness, which will be further confirmed in the experimental section.

Dataset construction

The effectiveness of learning-based phase retrieval and unwrapping depends on the quality and diversity of the training dataset. To capture the intricate complexities of real-world applications, we prioritize real-world data acquisition over synthetic alternatives. Our dataset encompasses a wide array of objects, including those with smooth surfaces (e.g., plastics and matte ceramics with simple geometries), moderately rough materials (e.g., wood and painted metal), and highly textured or reflective surfaces (e.g., intricately textured plaster sculptures and polished metals). This diverse selection enables the network to learn and generalize across both low-frequency structures and high-frequency features, which is essential for accurate phase retrieval and unwrapping under a wide range of conditions. Samples were collected under varying illumination conditions, including direct and diffuse ambient lighting, to simulate diverse environments. Each sample set comprises single-shot composite images ($I^{c}_{tri-fri}$) alongside additional reference images essential for ground-truth generation. As depicted in Fig. 3A, ground-truth phase values are derived from an N-step phase-shifting algorithm, ensuring high-precision reference data for effective

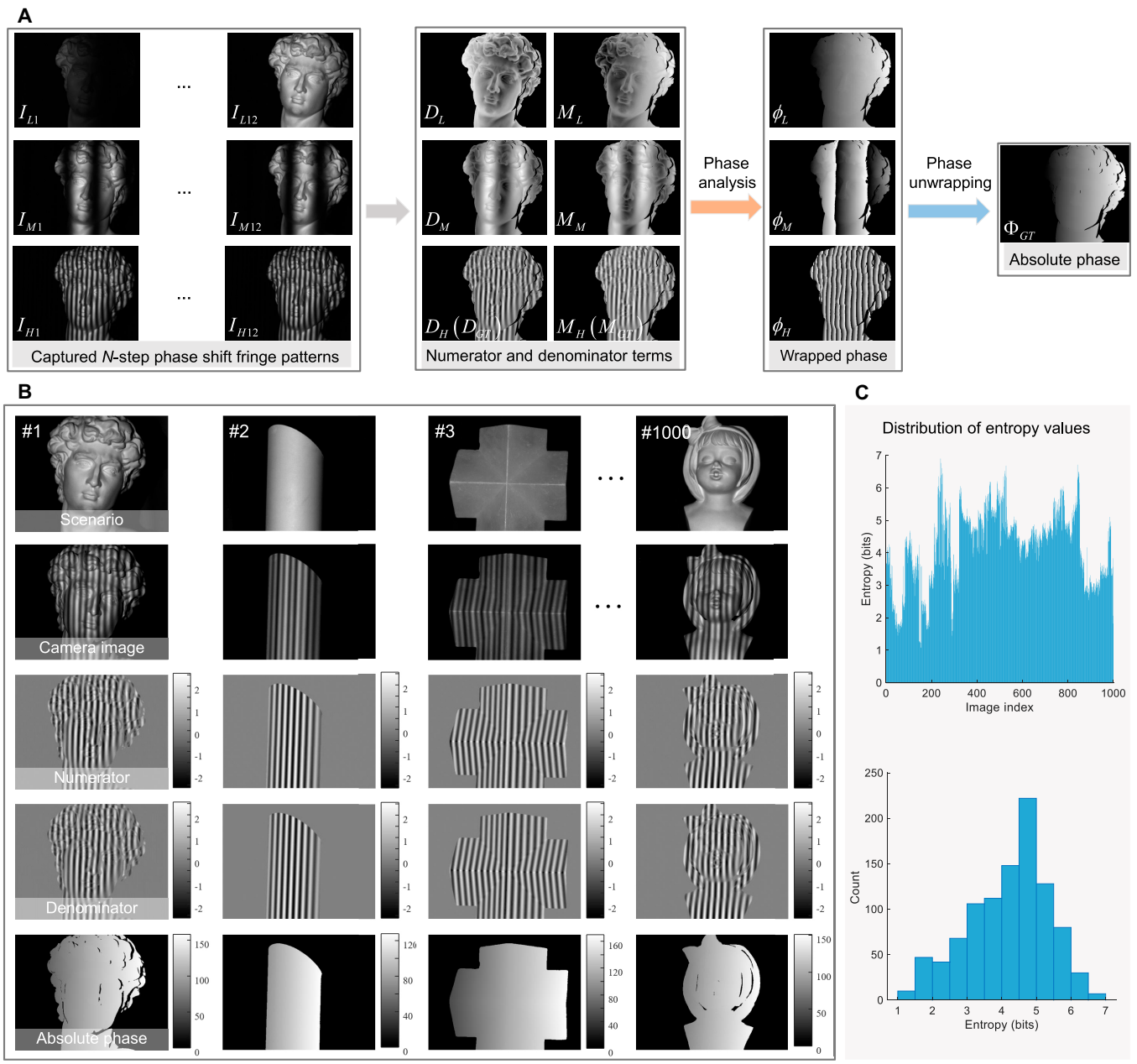

Fig. 3. Triangular-wave-embedded fringe image dataset preparation and analysis. (A) Workflow for generating training data: computation of the numerator (M_{GT}) and denominator (D_{GT}) terms of the wrapped phase from 12-step phase-shifted fringe images at 3 distinct frequencies, followed by phase unwrapping to obtain the absolute phase (Φ_{GT}). (B) Representative examples from the dataset, showcasing a variety of typical scenes with diverse surface textures and illumination conditions. (C) Entropy distribution of the dataset, illustrating a broad range from approximately 1 bit to 7 bits. The distribution features a primary peak around 4 bits, with significant clusters in the 2–3 bit and 5–6 bit ranges, reflecting the dataset's diversity in scene complexity.

training. Our comprehensive dataset (partially shown in Fig. 3B) includes composite images with triangular-wave-embedded fringes ($I^c_{tri-fri}$), the M_H, D_H terms corresponding to the phase of the high-frequency sinusoidal fringes (i.e., M_{GT} and D_{GT}), as well as its absolute phase (Φ_{GT}). In total, we compiled 1,000 distinct scene configurations, each captured from multiple viewpoints or object poses, resulting in 4,000 labeled images for training. This extensive coverage mitigates overfitting and enhances the model's generalization to real-world conditions.

To quantitatively assess the diversity of our dataset, we analyzed its entropy distribution. As illustrated in Fig. 3C, the entropy values span from approximately 1 bit to 7 bits, creating a broad overall range. The main peak lies near 4 bits, with notable clusters in the 2–3 bit and 5–6 bit ranges, making the distribution more dispersed. The prevalence of low-entropy (1 to 2 bits) samples reflects highly uniform or simple scenes, whereas values close to 7 bits indicate richer textures or higher noise levels. By covering both low-entropy (<2 bits) and high-entropy (>5 bits) extremes, the dataset provides robust scope for tasks requiring strong generalization, ensuring adaptability to widely varying scene complexities. Such a wide distribution is particularly advantageous if the model must handle extremely simple (near-uniform) or highly complex (dense noise/texture) scenarios in practical deployments, although training with a broad range may require additional strategies (e.g., careful hyperparameter tuning) to maintain performance across mid-range scenes. Through this balanced real-world dataset construction, we capture a wide range of surface textures, lighting conditions, and geometries essential for phase retrieval and unwrapping. By incorporating such diversity, our enhanced Swin-Unet benefits from varied features and scenarios, ultimately leading to improved robustness and accuracy in challenging 3D measurement tasks.

Experimental setup

We constructed a monocular FPP system consisting of a high-resolution digital light projector (LightCrafter 4500, 912 × 1,140 pixels) and a high-speed monochrome industrial camera (Basler acA640-750um, 640 × 480 pixels), as illustrated in Fig. 4A. The projector encodes and projects an 8-bit pattern that embeds low-frequency triangular waves within high-frequency sinusoidal fringes to illuminate the target surfaces. An industrial camera is triggered in sync with the projector, allowing for rapid capture of deformed fringe images while ensuring accurate imaging even in dynamic scenes. To minimize environmental light interference, the system operates in a controlled environment, ensuring optimal conditions for high-precision 3D measurement of complex surfaces, even under dynamic conditions. The system's camera and projector were calibrated using a circular calibration board to determine their intrinsic and extrinsic parameters [37,38], covering a measurement range of 200 × 200 × 100 mm.

The enhanced Swin-Unet model was trained on a PC with an NVIDIA 3080 10GB GPU using the PyTorch 1.7.1 framework in Python 3.7, accelerated by CUDA 11.0. The designed composite pattern is generated by embedding a triangular wave into a sinusoidal fringe. Although a higher sinusoidal frequency offers finer 3D detail, it also makes phase unwrapping more prone to error. Based on our projector's lateral resolution and practical experiments, we chose 19 pixels as the sinusoidal fringe wavelength. In order to satisfy the constraints in Eq. (4), we selected 51 pixels as the wavelength of the triangular wave, ensuring unambiguous unwrapping of the high-frequency wrapped phase. Each captured triangular-wave-embedded fringe images were normalized to pixel values ranging from 0 to 1 and then fed into the network. The output included the numerator and denominator terms, as well as the coarse absolute phase, which were compared against the ground truth to calculate the error. Training parameters included 500 epochs, a batch size of 2, and a learning rate of $3e^{-4}$. The AdamW optimizer was used to minimize the loss and iteratively optimize network parameters. Notably, Network 1 training required approximately 8.2 h per dataset, while Network 2 took about 7.1 h. Both networks were trained using a combined loss function:

$$\mathcal{L}_{\text{Loss}} = \lambda_{\text{MSE}} \cdot \text{MSE}\left(\hat{\phi}, \phi\right) + \lambda_{\text{SSIM}} \cdot \left[1 - \text{SSIM}\left(\hat{\phi}, \phi\right)\right] \quad (7)$$

where $\hat{\phi}$ represents the predicted phase term, ϕ denotes the ground truth, and λ_{MSE} and λ_{SSIM} are weighting factors for the mean squared error (MSE) and structural similarity index (SSIM) terms in the loss function.

Additionally, to evaluate the precision of our approach, we measured a standard component consisting of 2 ceramic spheres (Fig. 4B). The reconstructed surfaces of these spheres and their corresponding error distributions are displayed in Fig. 4C to E. Figure 4F and G shows that the root mean square (RMS) error values of the standard spheres are 0.060 and 0.064 mm, respectively. These low RMS values further indicate that our 3D reconstruction results have high accuracy, confirming the reliability of the method in practical applications.

Results

Assessing phase prediction accuracy across various coding strategies

This section outlines experiments conducted to validate the benefits of our multiplexed coding strategy and network architecture. First, we compare the accuracy of different coding strategies for single-frame phase prediction using DL by evaluating 3 scenes with distinct characteristics, as illustrated in Fig. 5. These include a simple polyhedral surface, a complex plaster statue, and 2 isolated, discontinuous plaster statues. Specifically, Fig. 5 (A1 to A3, B1 to B3, and C1 to C3) shows the captured single-frequency fringe images, dual-frequency composite fringe images, and triangular-wave-embedded fringe images for each scenario. Figure 5 (A4 to A6, B4 to B6, and C1 to C3) displays the absolute phase errors obtained using single-frequency fringe coding, dual-frequency composite fringe coding, and triangular-wave-embedded fringe coding strategies, respectively. Furthermore, Fig. 5 (A7 to A9, B7 to B9, and C7 to C9) illustrates the 3D results generated by each coding strategy, while Fig. 5 (A10, B10, and C10) represents the ground truth for the 3 scenes.

Single-frame measurement accuracy for different structured light coding strategies is evaluated using mean absolute error (MAE) between the predicted and true phase, quantifying the deviation of the network prediction from ground truth. A lower MAE indicates that the network's predictions are closer to the actual phase, demonstrating a higher accuracy of the network model on the test dataset. Additionally, Table 1 provides quantitative analysis data for the 3 scenarios, including the MAE of wrapped and unwrapped phases, as well as fringe order accuracy.

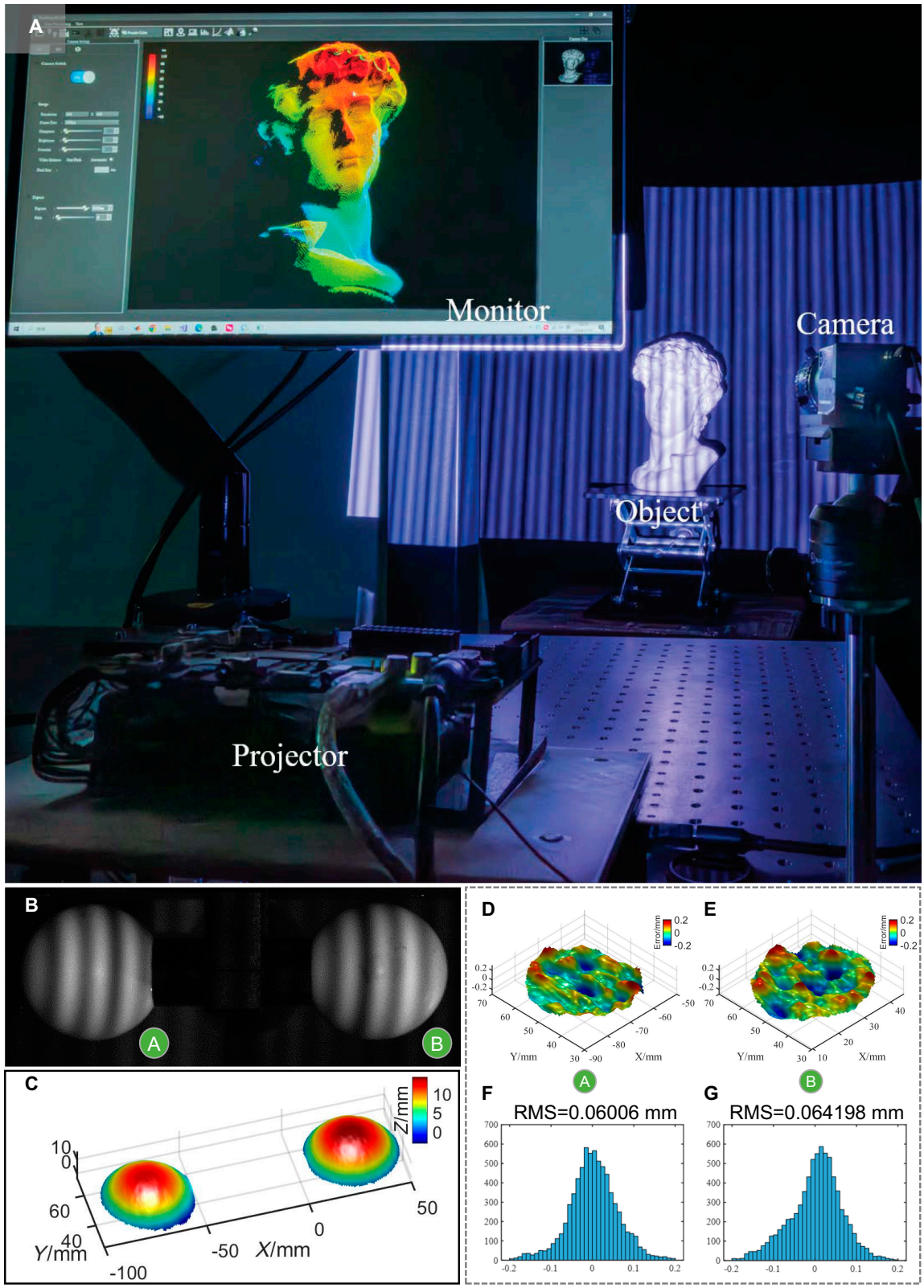

Fig. 4. Experimental setup and results for single-shot triangular-wave-embedded fringe projection 3D measurement. (A) Photograph of the constructed FPP system. The prototype setup contains a high-speed projector and industrial camera, complemented by a high-resolution monitor to display the reconstructed 3D point cloud results. (B) Captured multiplexed structured light image of 2 standard ceramic spheres (labeled A and B), used for precision analysis. (C) 3D surface reconstruction of the 2 spheres using the proposed method, with the color scale representing depth in millimeters. (D and E) Error distribution maps for the reconstructed surfaces of spheres A and B. (F and G) Histograms of the RMS error values for spheres A and B.

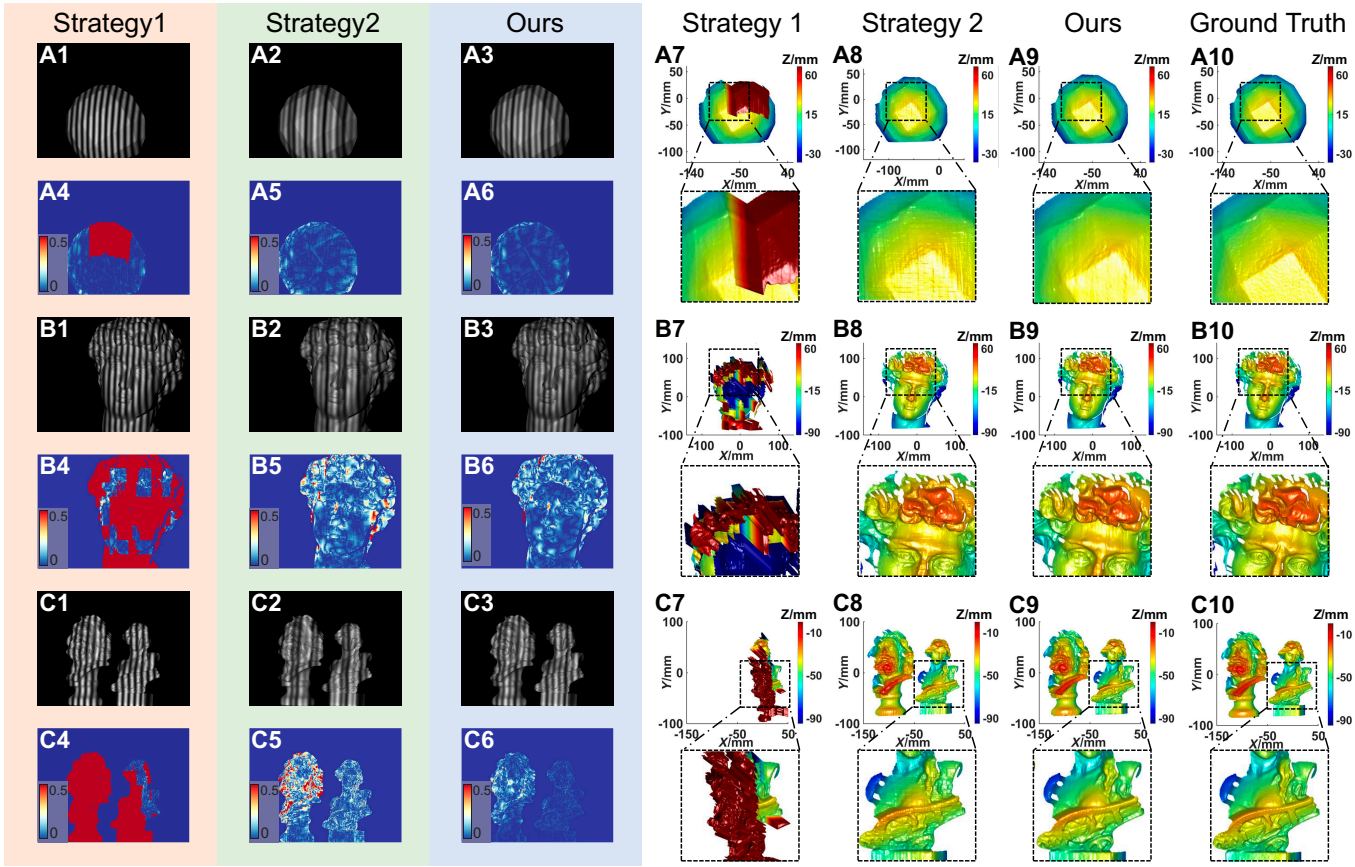

Fig. 5. Prediction results under different fringe coding strategies. (A1 to A3, B1 to B3, and C1 to C3) The structured light images were captured by sequentially projecting single-frequency fringes, dual-frequency composite fringes, and triangular-wave-embedded fringes (the proposed) across 3 scenarios. (A4 to A6, B4 to B6, and C4 to C6) Corresponding absolute phase results for 3 scenarios. (A7 to A9, B7 to B9, and C7 to C9) 3D reconstruction results from the single-frequency fringes, composite dual-frequency fringes, and triangular-wave-embedded fringe coding strategies, respectively, for the same 3 scenarios. (A10, B10, and C10) Ground-truth 3D reconstruction results for each scenario.

Table 1. Prediction results under different structured light coding strategies

Structured light coding strategy	Scene	Wrapped phase MAE (rad)	Accuracy (%)	Unwrapped phase MAE (rad)
Single-frequency fringe coding strategy	(a)	0.0168	71.724	1.5277
	(b)	0.0438	26.904	7.3421
	(c)	0.0334	12.225	25.0148
Dual-frequency composite fringe coding strategy	(a)	0.0319	99.628	0.0307
	(b)	0.1131	98.633	0.1064
	(c)	0.0423	99.424	0.0516
Triangular-wave-embedded fringe coding strategy	(a)	0.0230	99.754	0.0221
	(b)	0.0757	99.034	0.0774
	(c)	0.0343	99.543	0.0324

The results indicate that the phase distribution for single-frequency fringe images exhibits significant prediction errors, whether on simple polyhedral surfaces or between isolated objects. When deriving fringe order from single-frequency images using a neural network, the lack of additional auxiliary information hinders resolving fringe period ambiguities, resulting in poor reconstruction quality. In contrast, both the dual-frequency composite fringe coding strategy and the triangular-wave-embedded sinusoidal fringe coding strategy successfully achieve high-accuracy phase distribution and overcome fringe ambiguity using single-frame DL. However, the dual-frequency sinusoidal fringe coding strategy experiences a reduction in fringe contrast (see Fig. 5C1 and C2) due to the influence of the relatively low-frequency sinusoidal auxiliary signal on the high-frequency sinusoidal fringe to be demodulated, leading to a slight decrease in phase prediction

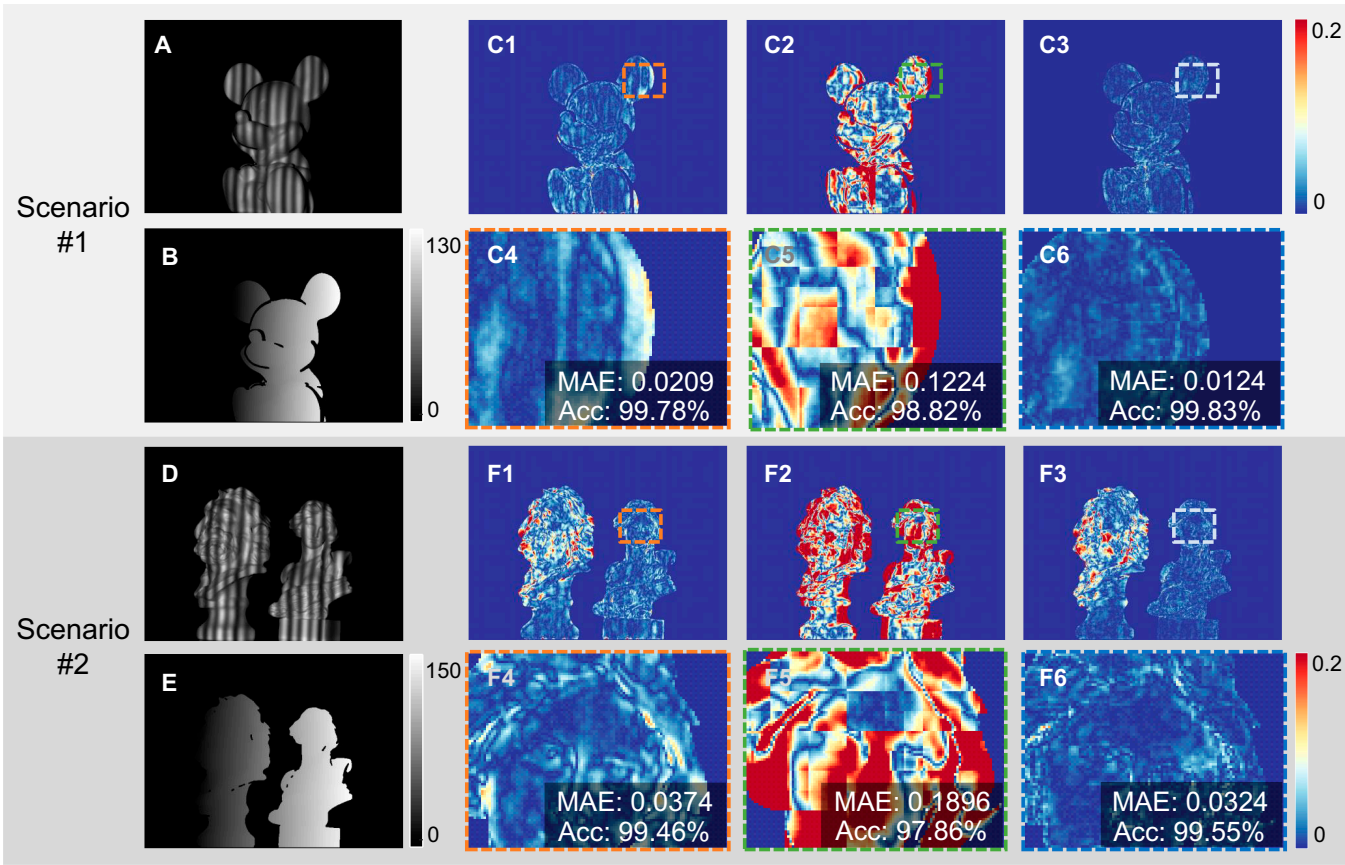

Fig. 6. Comparison of absolute phase predictions and error distributions from different networks. (A and D) Original triangular-wave-embedded fringe images from 2 distinct measurement scenarios. (B and E) Absolute phase results predicted by the enhanced Swin-Unet network. (C1 and F1) Absolute phase error distribution of the Unet network. (C4 and F4) Zoomed-in view of the orange box area in (C1) and (F1). (C2 and F2) Absolute phase error distribution of the original Swin-Unet network. (C5 and F5) Zoomed-in view of the green box area in (C2) and (F2). (C3 and F3) Absolute phase error distribution of the enhanced Swin-Unet network (the proposed), showing improved phase prediction accuracy. (C6 and F6) Zoomed-in view of the blue box area in (C3) and (F3).

accuracy. In comparison, the triangular-wave-embedded sinusoidal fringe composite coding strategy achieves higher precision in phase analysis and unwrapping, proving more effective than the dual-frequency sinusoidal strategy in resolving phase ambiguities.

Accuracy evaluation: Transformer versus CNN architectures

To evaluate the impact of the enhanced Swin-Unet's self-attention mechanism, we compared its phase prediction accuracy against the standard Unet and the original Swin-Unet. Both networks were configured with similar parameter counts and trained on the same dataset under identical hyperparameter settings as previously described. Using the triangular-wave-embedded fringe composite images as input, the 3 networks (UNet, original Swin-Unet, and enhanced Swin-Unet) were used to predict phase values. In Fig. 6A and D, we show the original triangular-wave-embedded composite images from 2 different measurement scenarios. Figure 6B and E depicts the absolute phase results predicted by our enhanced Swin-Unet network. The absolute phase error distribution maps for the Unet network are illustrated in Fig. 6C1 and F1, along with zoomed-in views of their corresponding regions (Fig. 6C4 and F4). Similarly, Fig. 6C2 and F2 presents the absolute phase error distributions for the original Swin-Unet network, while Fig. 6C3 and F3 showcases the error distributions for our enhanced Swin-Unet network. The results indicate that the enhanced Swin-Unet network achieves significantly higher absolute phase prediction accuracy compared to the original Swin-Unet network, regardless of whether the scene involves a single plaster model or a combination of 2 isolated models with discontinuous surfaces. The improved self-attention mechanism, with a larger attention window, enables the model to better capture the global contextual information within images, thereby facilitating a deeper understanding of image features. Unlike CNNs like Unet, which only have a local receptive field, the enhanced Swin-Unet network can also comprehend the global structure of fringe images and establish a correlation matrix among all pixel points in the image. This capability leads to the attainment of higher accuracy in the numerator, denominator, and absolute phase predictions.

In addition, the experimental results further validate the theoretical advantages of our data-driven approach, demonstrating that it significantly outperforms traditional phase recovery methods in mitigating speckle noise. By leveraging the prior knowledge embedded in deep neural networks and advanced self-attention mechanism, our method achieves robust and accurate phase recovery even under high noise conditions. This confirms the effectiveness of our approach for dynamic 3D shape measurement of objects with rough surfaces. As illustrated in Fig. 7, a comparison of 3D reconstruction results with and without

speckle noise is provided. Figure 7A to D shows the original fringe pattern, the reconstruction results obtained using the proposed Transformer-based enhanced Swin-Unet, the results from a traditional method, and corresponding zoomed-in views, respectively. Similarly, Fig. 7E to H depict the fringe pattern with added speckle noise, the reconstruction results using our DL method under high speckle noise, the results from the conventional method under the same noise conditions, and their zoomed-in views. These comparisons clearly highlight the superior performance of our method in noisy environments.

Application of dynamic industrial measurement

To validate the practical feasibility of our approach, we performed 3D measurements in 2 dynamic scenarios: a plaster model and an industrial workpiece. During the measurement process, the plaster model was placed on a controlled precision turntable to ensure stable and consistent rotation, which allowed for accurate data acquisition throughout the process. Figure 8A and B shows the camera images and reconstructed 3D results for the 12th and 75th frames, respectively, effectively capturing the detailed surface geometry. By employing a

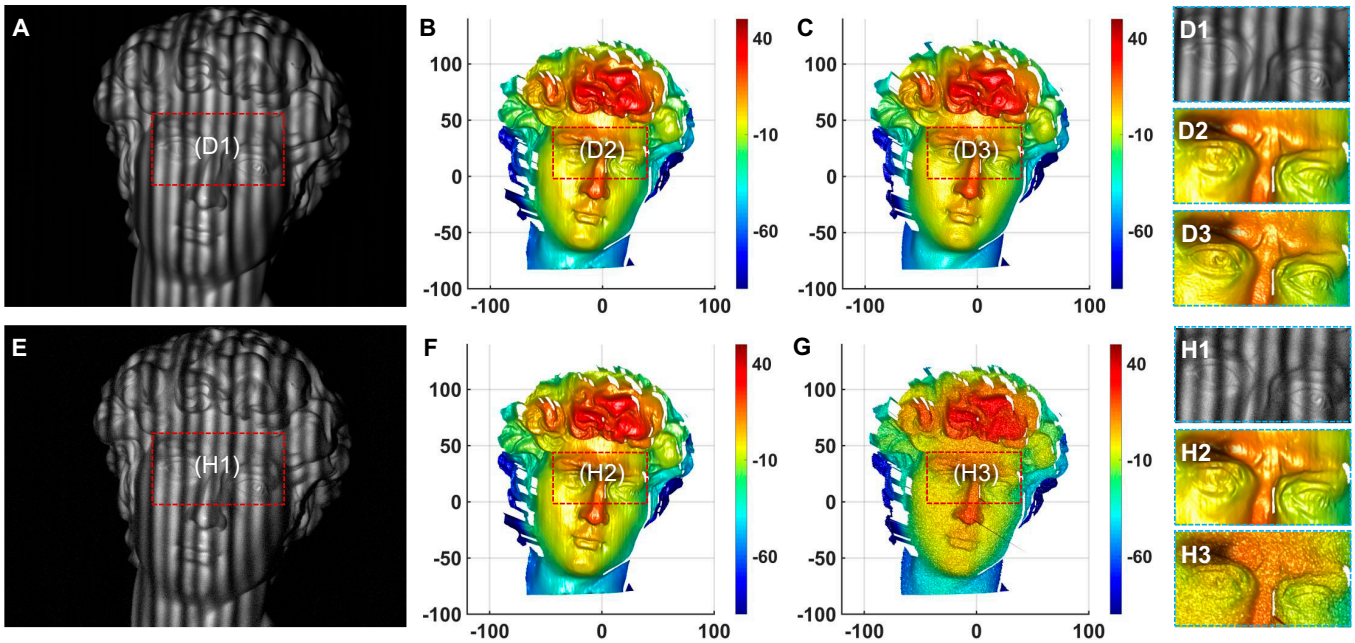

Fig. 7. Comparison of 3D reconstruction results with and without speckle noise. (A) Original fringe pattern. (B) Reconstruction results using the proposed Transformer-based enhanced Swin-Unet. (C) Traditional method results. (D1 to D3) Zoomed-in views of (A), (B), and (C). (E) Fringe pattern with added speckle noise. (F) Reconstruction results using the proposed deep learning method under high speckle noise. (G) Results from the conventional method under the same noise conditions. (H1 to H3) Zoomed-in views of (E), (F), and (G).

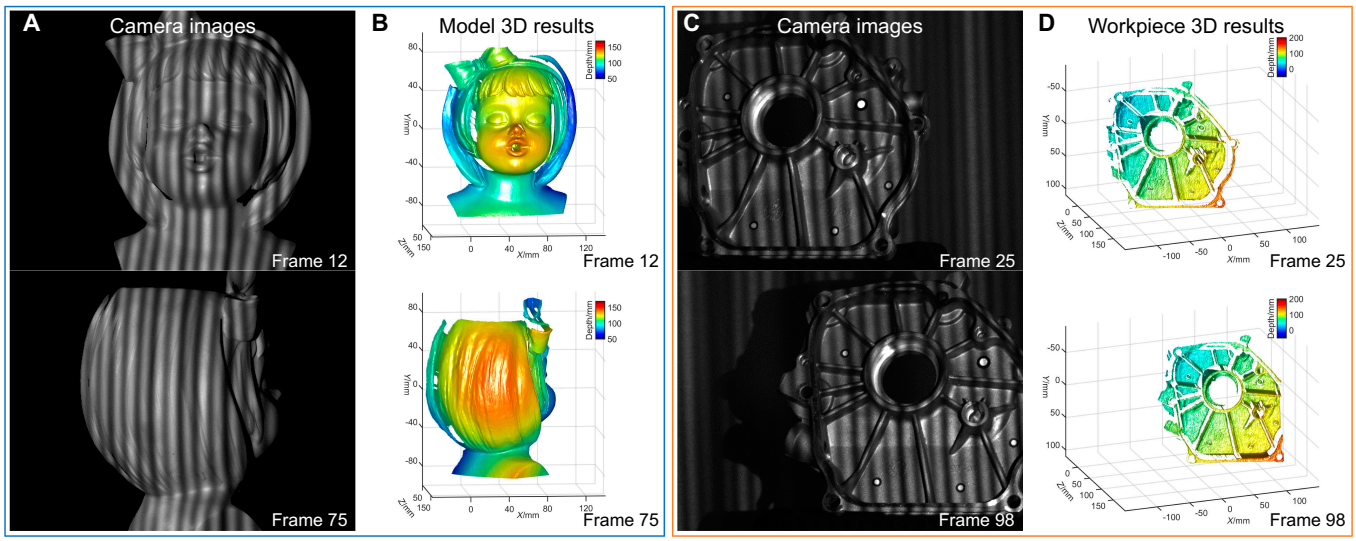

Fig. 8. 3D measurement results for complex dynamic scenes. (A and B) Camera images and 3D reconstruction results for the rotating plaster model at the 12th and 75th frames. (C and D) Images and reconstruction results for a moving industrial workpiece. Movies S1 and S2 further showcase the complete dynamic measurement processes for the plaster model and industrial workpiece, respectively.

single-frame reconstruction framework, our method inherently avoids motion-induced artifacts. Regardless of horizontal, vertical, or complex motion, our design maintains stable and precise 3D shape measurements. Furthermore, the high-frame-rate camera integrated into our system reliably captures fringe images even under rapid object motion, preserving reconstruction quality despite high-speed or complex movements. This approach substantially mitigates typical issues such as motion blur and temporal aliasing, thereby enhancing performance in dynamic scenarios.

Figure 8C and D illustrates a moving industrial workpiece. Despite changes in position and orientation, the system maintained high measurement accuracy, demonstrating robustness that is critical for quality control and inspection applications in manufacturing processes. Supplementary Materials include Movies S1 and S2, which showcase the dynamic measurement results for both the plaster model and the industrial workpiece, further illustrating the method's efficacy. Overall, these results highlight the feasibility of our DL-based approach for dynamic industrial detection and underscore its potential for precise real-time measurements in industrial manufacturing.

Conclusion

In this work, we introduce a single-frame 3D measurement approach that integrates triangular-wave-embedded fringe projection with Transformer-based networks, providing motion-immune, high-resolution reconstruction. By embedding a triangular wave into high-frequency sinusoidal fringes, our method preserves the integrity of the demodulated phase while minimizing interference in the high-frequency components. The resulting composite fringe image, used as the network input, significantly enhances phase retrieval and unwrapping, leading to more robust absolute phase recovery and extending the applicability of our method across diverse surface types and complex environments. Our experiments confirm that this approach outperforms traditional techniques, particularly in terms of phase accuracy and 3D reconstruction quality. Furthermore, the Transformer-based architecture strengthens the extraction of relevant fringe features, thereby improving both accuracy and robustness in phase unwrapping predictions.

Our method relies on the careful selection of triangular and sinusoidal frequencies, which may constrain flexibility under specific hardware or environmental conditions. Future work may involve optimizing the network's computational footprint (e.g., reducing model complexity or leveraging hardware accelerators) to enable real-time 3D imaging in high-speed industrial settings. Additionally, further research could explore applying this framework to dynamic or deformable surfaces, thereby expanding its utility in medical imaging, robotics, and industrial inspection.

Acknowledgments

Funding: This work was supported by the National Key Research and Development Program of China (2022YFB2804603), the National Natural Science Foundation of China (62075096 and U21B2033), the Leading Technology of Jiangsu Basic Research Plan (BK20192003), the "333 Engineering" Research Project of Jiangsu Province (BRA2016407), the Jiangsu Provincial "One belt and one road" innovation cooperation project (BZ2020007), the Fundamental Research Funds for the Central Universities (30919011222, 30920032101, 30921011208, and 2023102001), the China Postdoctoral Science Fund (2023T160318), and the Open Research Fund of Jiangsu Key Laboratory of Spectral Imaging & Intelligent Sense (JSGP202105 and JSGP202201).
Author contributions: Y.L. conducted all experiments, organized the figures, and wrote the main manuscript text. Y.X. contributed to the section on network architecture. J.Q. conceived the idea and refined the manuscript. C.Z. supervised the work. S.F., Q.C., and C.Z. provided funding support. All authors reviewed and revised the manuscript.
Competing interests: The authors declare that they have no competing interests.

Data Availability

Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

Supplementary Materials

Movies S1 and S2

References

1. Gåsvik KJ. *Optical metrology*. 3rd ed. Chichester: John Wiley & Sons; 2003.
2. Geng J. Structured-light 3D surface imaging: A tutorial. *Adv Opt Photon*. 2011;3(2):128–160.
3. Su X, Zhang Q. Dynamic 3-D shape measurement method: A review. *Opt Lasers Eng*. 2010;48(2):191–204.
4. Zuo C, Feng S, Huang L, Tao T, Yin W, Chen Q. Phase shifting algorithms for fringe projection profilometry: A review. *Opt Lasers Eng*. 2018;109:23–59.
5. Wu Z, Wang H, Chen F, Li X, Chen Z, Zhang Q. Dynamic 3D shape reconstruction under complex reflection and transmission conditions using multi-scale parallel single-pixel imaging. *Light Adv Manuf*. 2024;5:34.
6. Zuo C, Huang L, Zhang M, Chen Q, Asundi A. Temporal phase unwrapping algorithms for fringe projection profilometry: A comparative review. *Opt Lasers Eng*. 2016;85:84–103.
7. Heist S, Kühmstedt P, Tünnermann A, Notni G. Theoretical considerations on aperiodic sinusoidal fringes in comparison to phase-shifted sinusoidal fringes for high-speed three-dimensional shape measurement. *Appl Opt*. 2015;54(35):10541–10551.
8. Heist S, Lutzke P, Schmidt I, Dietrich D, Kuhmstedt P, Tunnerman A, Notni G. High-speed three-dimensional shape measurement using GOBO projection. *Opt Lasers Eng*. 2016;87:90–96.
9. Pan B, Xie H, Wang Z, Qian K, Wang Z. Study on subset size selection in digital image correlation for speckle patterns. *Opt Express*. 2008;16(10):7037–7048.
10. Zhang Z. Review of single-shot 3D shape measurement by phase calculation-based fringe projection techniques. *Opt Lasers Eng*. 2012;50(8):1097–1106.
11. Gao L, Liang J, Li C, Wang LV. Single-shot compressed ultrafast photography at one hundred billion frames per second. *Nature*. 2014;516:74–77.
12. Jing X, Zhao R, Li X, Jiang Q, Li C, Geng G, Li J, Wang Y, Huang L. Single-shot 3D imaging with point cloud projection based on metadevice. *Nat Commun*. 2022;13:7842.

13. Takeda M, Mutoh K. Fourier transform profilometry for the automatic measurement of 3-D object shapes. *Appl Opt*. 1983;22:3977–3982.
14. Kemao Q. Two-dimensional windowed Fourier transform for fringe pattern analysis: Principles, applications and implementations. *Opt Lasers Eng*. 2007;45(2):304–317.
15. Barbastathis G, Ozcan A, Situ G. On the use of deep learning for computational imaging. *Optica*. 2019;6(8):921–943.
16. Lin X, Rivenson Y, Yardimci NT, Veli M, Luo Y, Jarrahi M, Ozcan A. All-optical machine learning using diffractive deep neural networks. *Science*. 2018;361(6406):1004–1008.
17. Zuo C, Qian J, Feng S, Yin W, Li Y, Fan P, Han J, Qian K, Chen Q. Deep learning in optical metrology: A review. *Light Sci Appl*. 2022;11(1):39.
18. Wang K, Song L, Wang C, Ren Z, Zhao G, Dou J, Di J, Barbastathis G, Zhou R, Zhao J, et al. On the use of deep learning for phase recovery. *Light Sci Appl*. 2024;13(1):4.
19. Feng S, Chen Q, Gu G, Tao T, Zhang L, Hu Y, Yin W, Zuo C. Fringe pattern analysis using deep learning. *Adv Photonics*. 2019;1:25001.
20. Van der Jeught S, Dirckx JJ. Deep neural networks for single shot structured light profilometry. *Opt Express*. 2019;27(12):17091–17101.
21. Shi J, Zhu X, Wang H, Song L, Guo Q. Label enhanced and patch based deep learning for phase retrieval from single frame fringe pattern in fringe projection 3D measurement. *Opt Express*. 2019;27(20):28929–28943.
22. Nguyen AH, Ly KL, Li CQ, Wang Z. Single-shot 3D shape acquisition using a learning based structured-light technique. *Appl Opt*. 2022;61(29):8589–8599.
23. Wang K, Li Y, Kemao Q, Di J, Zhao J. One-step robust deep learning phase unwrapping. *Opt Express*. 2019;27(10):15100–15115.
24. Takeda M, Gu Q, Kinoshita M, Takai H, Takahashi Y. Frequency-multiplex Fourier transform profilometry: A single-shot three-dimensional shape measurement of objects with large height discontinuities and/or surface isolations. *Appl Opt*. 1997;36(22):5347–5354.
25. Wissmann P, Schmitt R, Forster F. Fast and accurate 3D scanning using coded phase shifting and high speed pattern projection. In: *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*. Hangzhou: IEEE; 2011. p. 108–115.
26. Wang Y, Zhang S. Novel phase-coding method for absolute phase retrieval. *Opt Lett*. 2012;37(11):2067–2069.
27. Tao T, Chen Q, Da J, Feng S, Hu Y, Zuo C. Real-time 3-D shape measurement with composite phase-shifting fringes and multi-view system. *Opt Express*. 2016;24(18):20253–20259.
28. Qian J, Feng S, Li Y, Tao T, Han J, Chen Q, Zuo C. Single-shot absolute 3D shape measurement with deep-learning based color fringe projection profilometry. *Opt Lett*. 2020;45(7):1842–1845.
29. Li Y, Qian J, Feng S, Chen Q, Zuo C. Deep-learning-enabled dual-frequency composite fringe projection profilometry for single-shot absolute 3D shape measurement. *Opto-Electron Adv*. 2022;5(5):210021.
30. Li Y, Qian J, Feng S, Chen Q, Zuo C. Composite fringe projection deep learning profilometry for single-shot absolute 3D shape measurement. *Opt Express*. 2022;30(3):3424–3442.
31. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Munich: Springer; 2015. p. 234–241.
32. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas (NV): IEEE; 2016. p. 770–778.
33. Ibtehaz N, Rahman MS. MultiResUNet: Rethinking the U-net architecture for multimodal biomedical image segmentation. *Neural Netw*. 2020;121:74–87.
34. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tang Y, Xiao A, Xu C, Xu Y, et al. A survey on vision transformer. *IEEE Trans Pattern Anal Mach Intell*. 2022;45(1):87–110.
35. Khan S, Naseer M, Hayat M, Zamir SW, Khan FS, Shah M. Transformers in vision: A survey. *ACM Comput Surv*. 2022;54(10):1–41.
36. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, Wang M. Swin-unet: Unet-like pure transformer for medical image segmentation. In: *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*. Israel: Springer; 2023. p. 205–218.
37. Zhang S, Huang PS. Novel method for structured light system calibration. *Opt Eng*. 2006;45(8):83601.
38. Feng S, Zuo C, Zhang L, Tao T, Hu Y, Yin W, Qian J, Chen Q. Calibration of fringe projection profilometry: A comparative review. *Opt Lasers Eng*. 2021;143:Article 106622.