Chapter 8
Digital Twin-Based Computer Vision Method for Robust Structural Vibration Estimation in the Underwater Environment

Enjian Cai, Zihan Wu, Michael D. Todd, and Zhen Hu

Abstract Structural Health Monitoring (SHM) of large civil infrastructures in inland waterways presents significant challenges, primarily because the critical components for damage assessment are often submerged. Even though sensors like strain gauges can provide valuable insights into structural dynamic response, their effectiveness is affected by the limited number of sensors and how the sensors are placed. To address these limitations, Computer Vision (CV)-based vibration estimation methods have been developed, offering full-field measurements and overcoming the spatial constraints of sparse sensors. However, the efficacy of CV-based approaches declines drastically in underwater environments due to poor visibility conditions, which significantly affect image quality and measurement reliability. Motivated by overcoming the challenge of underwater SHM using CV-based methods, this paper proposes a Digital Twin (DT)-based approach for Underwater Image Enhancement (UIE) to enable accurate sub-pixel vibration estimation. First, a DT-based Physics-Based Graphics Model (PBGM) is employed to generate sufficient datasets for the various water scenarios, addressing the scarcity of underwater images and ground-truth labels. Then, an Edge-Preserving Dual Attention Transformer (EDUT) with an edge-preserving loss is designed to eliminate image degradation while retaining fine structural details. Finally, the sub-pixel template matching method is applied for vibration estimation. The performance of the proposed method is demonstrated through a case study using the Greenup miter gate in the United States, achieving an average PSNR of 27.81 dB for UIE and 0.0485 to 0.2319 sub-pixel accuracy in vibration estimation.

Keywords Digital twin · Computer vision · Underwater image enhancement · Vibration estimation

Introduction

Structures like dams and bridges in inland waterways are essential infrastructure components. However, aging, deterioration, and climate change-induced flooding threaten structural integrity, making cost-effective underwater inspection crucial [1]. Current underwater inspection primarily uses human divers, sonar scanning, and Computer Vision (CV) techniques for static surface defect detection [2]. While CV integrated with Unmanned Underwater Vehicles (UUVs) offers automated and convenient capabilities and eliminates safety risks [3], these surface-based methods cannot reveal internal structural damage, which requires vibration-based identification to capture dynamic characteristics and modal properties [4].

Existing underwater structural vibration detection research is extremely limited [5]. Terrestrial methods using contact sensors face complex underwater installation and corrosion issues, providing only localized measurements [6]. LiDAR systems suffer from water absorption and scattering effects, limiting their underwater effectiveness. CV methods offer potential

Enjian Cai · Zhen Hu (✉)
Department of Industrial and Manufacturing Systems Engineering, University of Michigan-Dearborn,
4901 Evergreen Road, Dearborn, MI 48128, USA
e-mail: enjcai@umich.edu; zhennhu@umich.edu

Zihan Wu · Michael D. Todd
Department of Structural Engineering, University of California San Diego,
9500 Gilman Drive Mail Code 0085, La Jolla, CA 92093, USA
e-mail: z5wu@ucsd.edu; mdtodd@ucsd.edu

Kristin B. Zimmerman, PhD (eds.), *Model Validation and Uncertainty Quantification, 2026, Vol. 3*, Conference Proceedings of the Society for Experimental Mechanics Series, https://doi.org/10.1007/97887-438-1442-9_8

using UUVs with 2K+ cameras, but vibration identification requires sub-millimeter precision and sub-pixel measurement capabilities [7]. However, underwater images may be severely degraded by color casts, low contrast, and blurred details due to light effects. In this regard, existing Underwater Image Enhancement (UIE) methods face limitations: requiring extensive realistic datasets difficult to obtain in varying conditions [2], and relying on Cycle-Consistent Generative Adversarial Networks (CycleGAN) approaches with training instability, limited enhancement precision, and domain adaptation challenges [8].

To this end, this paper proposed a comprehensive Digital Twin (DT)-based underwater dataset generation method and an Edge-Preserving Dual Attention Transformer (EDUT) approach. Main contributions include: (1) enabling sub-pixel vibration measurement for underwater structural dynamic monitoring, (2) developing a Physics-Based Graphics Model (PBGM) creating the realistic three-dimensional (3D) underwater scenarios, and (3) introducing EDUT with Dual Attention Transformer Block (DATB) and edge-preserving loss function to address CycleGAN and Transformer limitations and preserve fine details [8, 9].

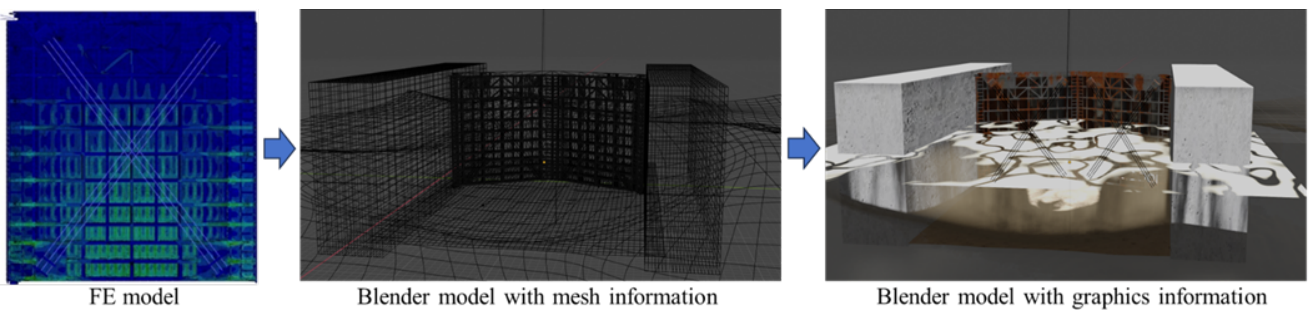

Fig. 1 Framework of DT-based underwater data generation

DT-Based Underwater Data Generation

The implementation of PBGM-based water-quality simulation follows a comprehensive methodology integrating structural modeling with aquatic environment reproduction. A Finite Element (FE) model of the underwater structure developed in Abaqus is first validated to capture structural behaviors, with geometric specifications input into Blender 4.0 using principled BSDF for material appearance. The framework, as shown in Figure 1, employs computational fluid-dynamics principles with physically accurate optical modeling to generate turbid water conditions based on the Greenup miter gate environment, representing sediment-laden waterways affected by erosion. The methodology involves four key stages. First, mesh validation uses Gerstner wave mathematics for surface topology. Second, the material setup applies warm base colors and volume scattering at 0.015 density. Third, lighting uses warm illumination at 4.0 energy units. Finally, turbid water simulation employs muddy substrates with sparse vegetation and 30 bubbles using increased drag coefficients (0.15) for sediment effects.

Edut Method for Effective UIE

The EDUT method architecture, as shown in Figure 2, employs a hierarchical encoder-decoder structure with three encoder layers and three decoder layers connected by skip connections for multi-scale feature processing. This design enables enhancement of objects at different underwater distances where scattering effects vary significantly. A bottleneck layer captures global pixel dependencies while maintaining computational efficiency for large-scale underwater structures.

EDUT addresses underwater degradation through three main components. Firstly, the Channel Self-Attention Transformer (CSAT) component handles wavelength-dependent color distortion by weighting channel importance, recognizing that red light penetrates only 3-4 m while blue-green light reaches deeper depths. This selective attention prevents single-channel dominance, crucial for detecting materials and corrosion in varying water conditions. Secondly, the Shifted Window Pixel Self-Attention Transformer (SW-PSAT) component addresses distance-dependent degradation through local window-based attention, enabling cross-window connections essential for continuous structural features like pipelines and hull sections that span multiple processing windows. The connection scheme and layers for CSAT and SW-PSAT are consistent with those in [10].

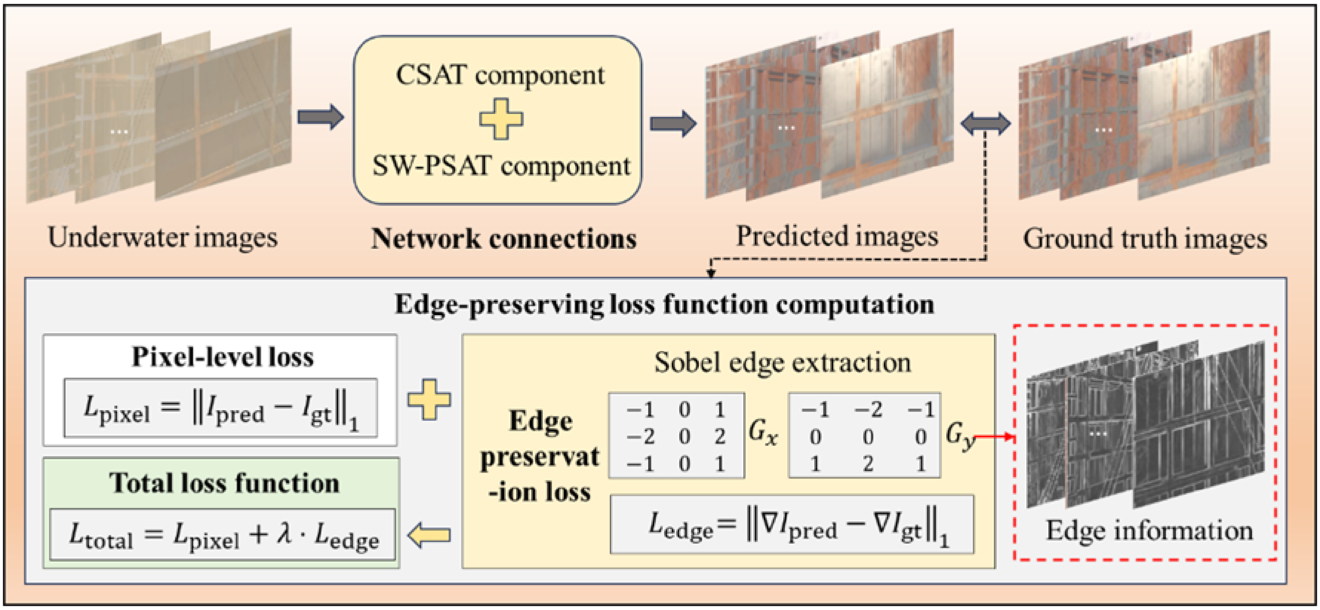

Fig. 2 Architecture of the EDUT method for effective UIE

Thirdly, the proposed edge-preserving loss function addresses the limitations of standard L_2 loss methods that treat all pixels equally, overlooking structural boundaries essential for underwater inspection. Underwater environments present unique degradation challenges: bubble occlusion creates false edges and obscures structural boundaries, scattering blur from suspended particles softens edges, reducing detection accuracy, and non-uniform degradation causes structural elements to appear with varying clarity depending on distance and water conditions.

To address these challenges, the proposed edge-preserving loss function incorporates both pixel-level reconstruction and structural boundary preservation. The pixel-level reconstruction loss $L_{\text{pixel}} = \|I_{\text{pred}} - I_{\text{gt}}\|_1$ uses L_1 norm for its robustness to outliers, while edge information is computed using the Sobel operator $\nabla I = \sqrt{(G_x * I)^2 + (G_y * I)^2}$ where G_x and G_y are horizontal and vertical Sobel kernels. The edge preservation loss $L_{\text{edge}} = \|\nabla I_{\text{pred}} - \nabla I_{\text{gt}}\|_1$ ensures the network maintains both edge strength and directional accuracy. The complete loss function $L_{\text{total}} = L_{\text{pixel}} + \lambda \cdot L_{\text{edge}}$ combines these components, where the weight parameter $\lambda = 0.2$ is carefully calibrated to balance edge preservation without introducing artificial sharpening that could be mistaken for structural defects. This formulation forces the EDUT network to maintain sharp, geometrically accurate structural boundaries while preserving fine details crucial for sub-pixel underwater vibration detection.

Sub-Pixel Underwater Structural Vibration Estimation

After EDUT enhancement, sub-pixel template matching is used to estimate structural vibrations from the enhanced video sequence, where the first frame serves as a reference, as shown in Figure 3. Mean Squares (MS) is employed as the similarity

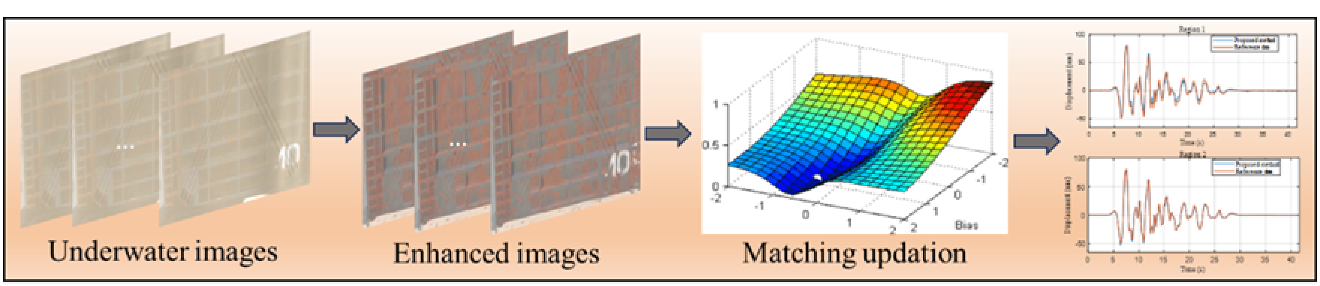

Fig. 3 Architecture of the sub-pixel template matching for underwater structural vibration estimation

metric for its superior noise resistance in intensity-based matching applications. The MS between reference frame R_0 and current frame R_i incorporates spatial displacement parameters to quantify structural movement. The similarity metric is computed as $\mathrm{MS}(u,v) = (1/N)\Sigma\Sigma\left[R_0(x,y) - R_{\mathrm{i}}(x+u,y+v)\right]^2$, where u and v represent horizontal and vertical structural vibrations in pixels, $R_0(x,y)$ and $R_i(x+u,y+v)$ are intensity values of corresponding structural regions, and N is the number of pixels in the overlapping area. This formulation captures template correlation across frames, accounting for translational displacement. Since structures primarily undergo rigid-body translation, optimization focuses only on translational parameters. The optimal structural vibrations with units of pixels are determined by minimizing the MS metric $(\hat{u},\hat{v}) = \arg\min \mathrm{MS}(u,v)$ to identify displacement values with sub-pixel accuracy. Physical structural vibrations are then obtained by multiplying pixel vibrations with calibration factors derived from known structural dimensions.

Case Study

The proposed method was validated using the Greenup miter gate on the Ohio River, Kentucky, United States. A high-fidelity graphical model was generated in Blender 4.0 from the validated FE model, with yellow turbid water conditions simulating the gate's actual environment, as already shown in Figure 1. Virtual camera trajectories were systematically planned to simulate UUV navigation, capturing both gate sides from multiple viewpoints with 3000 frames at 1080×1920 resolution and 24 Hz sampling frequency. Comprehensive comparisons were conducted with state-of-the-art approaches: enhanced Underwater (U)-GAN [11] and U-Shape Transformer (UST) [9] method. Additionally, the EDUT trained on the authoritative UIE Benchmark (UIEB) dataset [12] was compared to assess the advantages of the proposed dataset. All networks were trained for 120 epochs with identical settings, using 256×256 resolution during training. For vibration validation, structural excitations along the Z-axis were applied to both gates using the CNP196 component of the 1994 Northridge earthquake [13]. Testing videos comprised 1001 frames captured at the same resolution and sampling frequency, with the initial frames of the vibration videos and two measurement regions shown in Figure 4. Pixel-to-physical conversion factors were calculated as 9.0765 mm/pixel and 11.0869 mm/pixel for left and right gates, respectively. All experiments were conducted on a workstation with an Intel Xeon W-2125 CPU, 64 GB RAM, and dual NVIDIA GeForce RTX 2080 Ti GPUs.

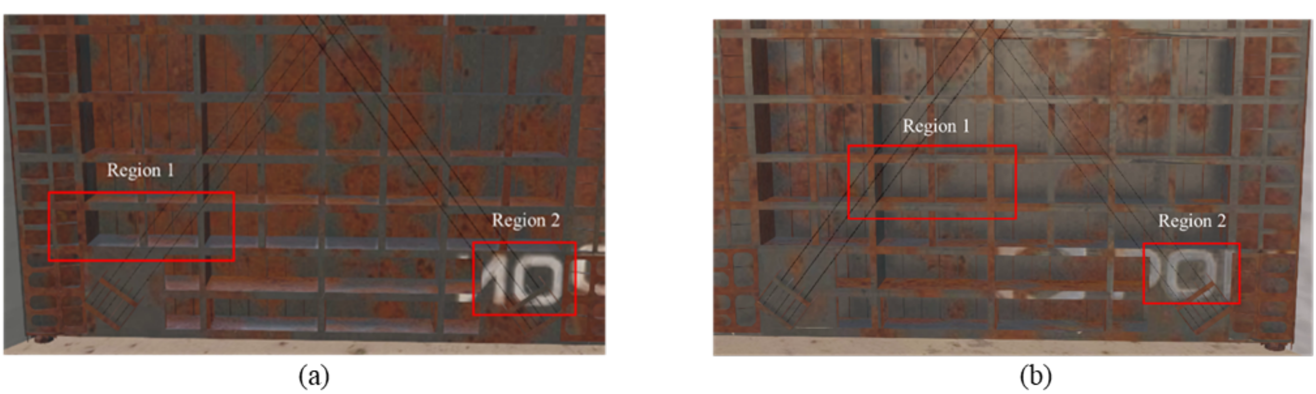

Fig. 4 Rendered frames recorded by two virtual cameras during the vibration at the 1st frame: (a) left gate and (b) right gate

After generating the yellow turbid 3D underwater environment, two virtual cameras were positioned along navigation routes to capture videos, producing 3000 frames per gate at a 24 Hz sampling frequency. The corresponding underwater videos and ground truth sequences are shown in Video 1. It could be observed that, with PBGM integration, the generated dataset realistically reflected variations in water color, bubble behavior, and lighting as the camera location and viewpoint changed. These images were used to train four models: DT-trained U-GAN, DT-trained UST, UIEB-trained EDUT, and DT-trained EDUT. The resulting models were applied to vibration videos of each gate in the yellow turbid environment, each containing 1001 frames. To save pages, only the left gate results are shown in Figure 5, with only the 500^{th} frame of enhanced images shown, while full sequences enhanced by DT-trained EDUT for both left and right gates are available in Video 2 and Video 3. As observed, DT-trained U-GAN and UST enhanced images were close to the ground truth but failed to fully recover gate textures. U-GAN results appeared blurred due to high-frequency detail smoothing to satisfy the discriminator, while UST outputs were overly bright because global brightness adjustment with patch-level attention reduced local control and amplified uneven illumination. UIEB-trained EDUT removed only a small portion of water effects due to dataset limitations. In contrast, DT-trained EDUT produced enhanced images closely matching ground truth with well-preserved gate edge textures, achieved through PBGM-based dataset generation and shifted-window pixel-level attention with edge-preserving loss. In addition, DT-trained EDUT reached the highest Peak Signal-to-Noise Ratio (PSNR) values, averaging 27.81 dB.

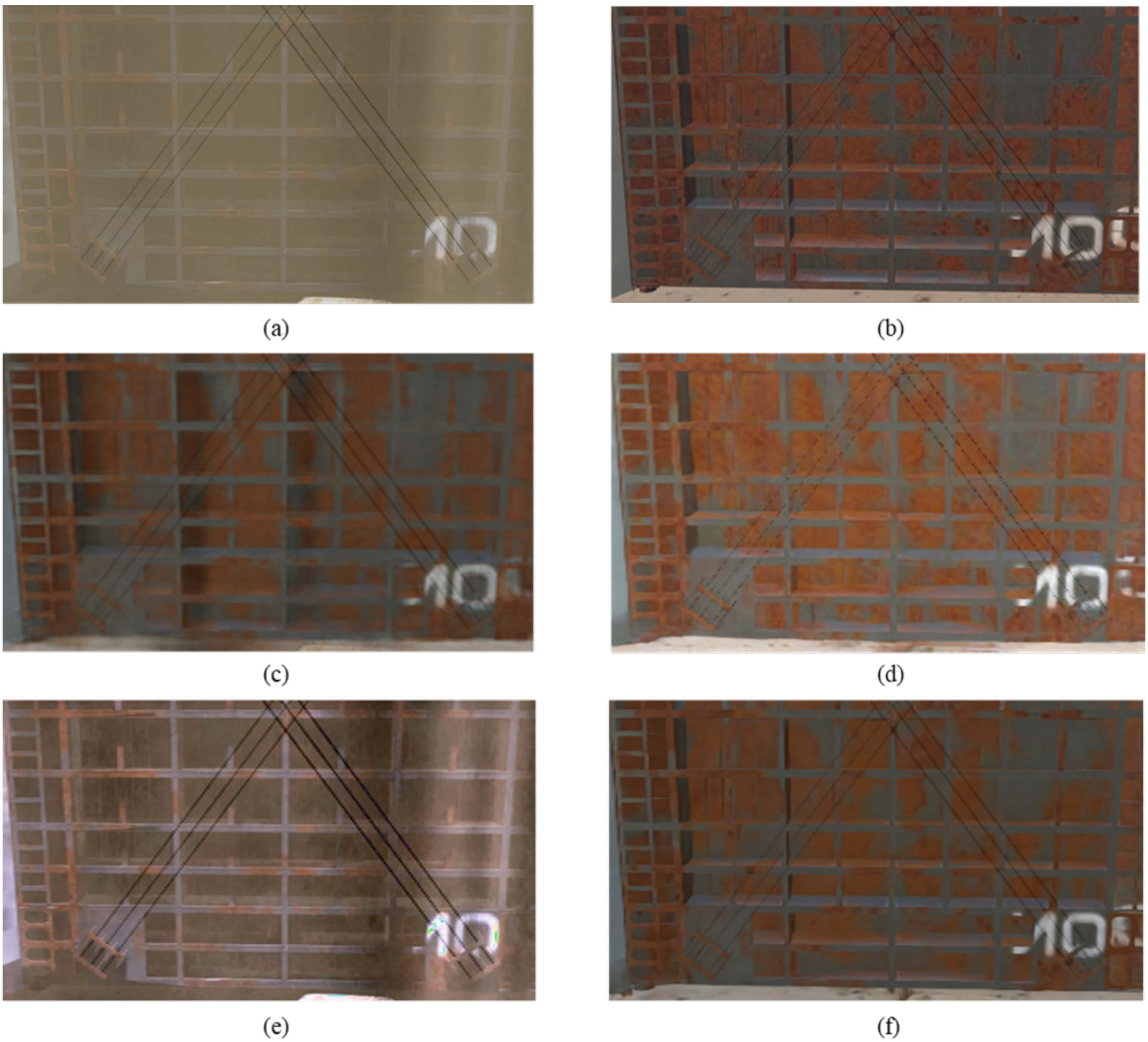

Fig. 5 The 500th recorded and enhanced frames from the 3D yellow turbid underwater environment when the left gate was vibrating: (a) raw underwater image, (b) ground truth image, and images enhanced by models of (c) DT-trained U-GAN, (d) DT-trained UST, (e) UIEB-trained EDUT, and (f) DT-trained EDUT

Then, sub-pixel template matching was applied to the enhanced videos to estimate structural vibrations, with resulting time-history curves shown in Figure 6, Video 4, and Video 5. Only videos enhanced by DT-trained EDUT were used for estimation due to superior performance. The estimated vibrations closely matched reference data, with Correlation Coefficients (CCs) above 0.9790 in all regions and reaching 0.9970 in Region 2. Mean Absolute Errors (MAEs) were 0.4400 mm and 0.8342 mm in Region 2, and 2.1326 mm and 2.5705 mm in Region 1, corresponding to sub-pixel accuracy of approximately 0.0485 to 0.2319 pixels based on conversion factors of 9.0765 mm/pixel and 11.0869 mm/pixel for left and right gates,

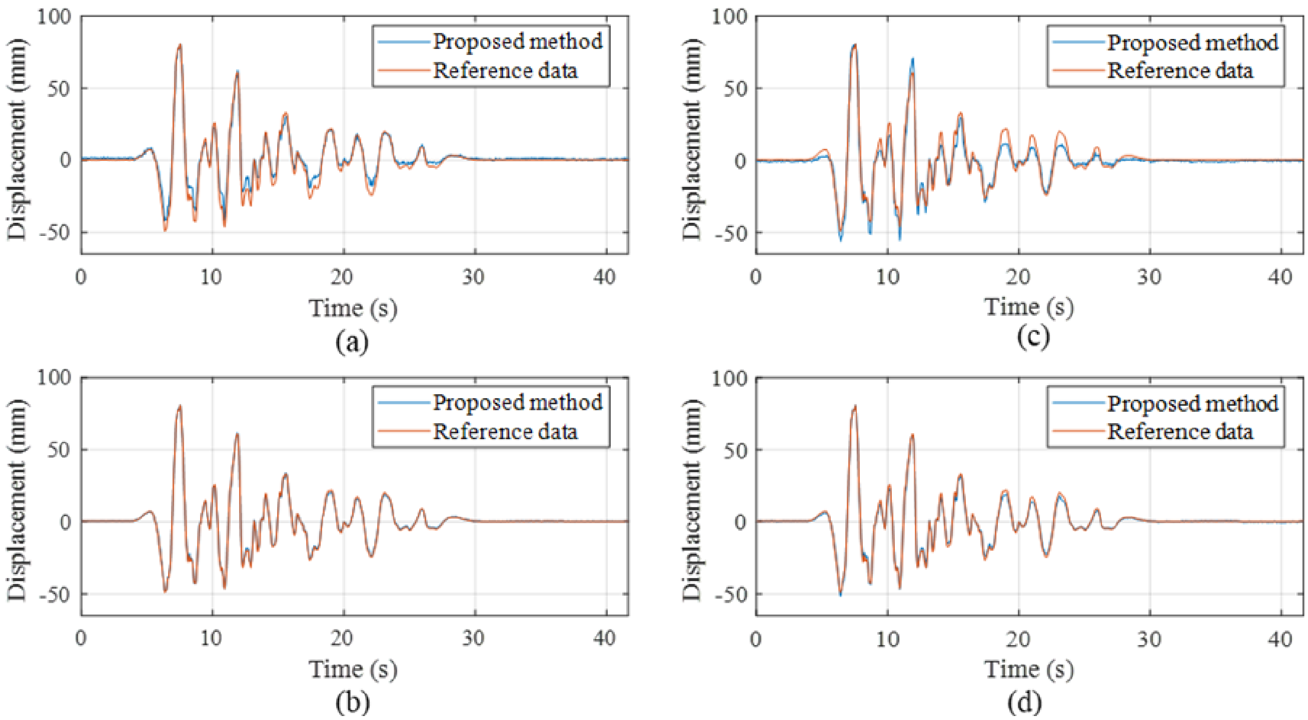

Fig. 6 Estimated structural vibrations from the 3D yellow turbid underwater environment in (a) Region 1, (b) Region 2 of the left gate, and (c) Region 1, (d) Region 2 of the right gate

respectively. The accuracy of estimated structural vibrations in Region 2 showed higher accuracy than that in Region 1, due to less complex textures that facilitated better UIE and reduced measurement artifacts.

Conclusions

In this paper, a DT-based method was proposed for effective underwater sub-pixel structural vibration estimation, by addressing severe image degradation in underwater scenes. A DT-based data generation method integrated with PBGM was proposed to generate sufficient underwater datasets for yellow turbid water conditions characteristic of sediment-laden rivers. Then, EDUT was proposed for effective UIE while preserving detailed edges, and sub-pixel template matching was applied to estimate structural vibrations with high precision. Through application on the Greenup miter gate under the yellow-turbid water condition, the proposed method could achieve an average PSNR of 27.81 dB for UIE with structural-vibration accuracy of approximately 0.0485 to 0.2319 sub-pixel. Future studies will extend the proposed method for out-of-plane vibration identification and elimination of camera motions from water surface disturbances.

Acknowledgments This work was supported by the US Army Corps of Engineers through the U.S. Army Engineer Research and Development Center Research Cooperative Agreement W9132T-22-20014. The support is gratefully acknowledged.

References

1. H. Tan, L. Zheng, C. Ma, Y. Xu, Y. Sun, Deep learning-assisted high-resolution sonar detection of local damage in underwater structures, Automation in Construction, 164 (2024) 105479.
2. S. Talamkhani, K. Liu, Underwater vision-enhanced image segmentation for supporting automated inspection of underwater bridge components, Automation in Construction, 175 (2025) 106230.
3. Z. Wu, Z. Hu, M.D. Todd, Uncertainty Quantification for Deep Learning–Based Automatic Crack Detection in the Underwater Environment, Springer Nature Switzerland, Cham, 2025, pp. 133-136.
4. E. Cai, Y. Zhang, X. Lu, P. Li, T. Zhao, G. Lin, W. Guo, A target-free video structural motion estimation method based on multi-path optimization, Mechanical Systems and Signal Processing, 198 (2023) 110452.

5. T. Chu, T. Nguyen, H. Yoo, J. Wang, A review of vibration analysis and its applications, Heliyon, 10 (2024) e26282.
6. J. Li, G. Chen, M.F. Antwi-Afari, Recognizing sitting activities of excavator operators using multi-sensor data fusion with machine learning and deep learning algorithms, Automation in Construction, 165 (2024) 105554.
7. E. Cai, Y. Zhang, S.T. Quek, Visualizing and quantifying small and nonstationary structural motions in video measurement, Computer-Aided Civil and Infrastructure Engineering, 38 (2023) 135-159.
8. Z. Lv, S. Dong, Z. Xia, J. He, J. Zhang, Enhanced real-time detection transformer (RT-DETR) for robotic inspection of underwater bridge pier cracks, Automation in Construction, 170 (2025) 105921.
9. L. Peng, C. Zhu, L. Bian, U-Shape Transformer for Underwater Image Enhancement, IEEE Transactions on Image Processing, 32 (2023) 3066-3079.
10. Z. Shen, H. Xu, T. Luo, Y. Song, Z. He, UDAformer: Underwater image enhancement based on dual attention transformer, Computers & Graphics, 111 (2023) 77-88.
11. C. Fabbri, M.J. Islam, J. Sattar, Enhancing Underwater Imagery Using Generative Adversarial Networks, 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 7159-7165.
12. C. Li, C. Guo, W. Ren, R. Cong, J. Hou, S. Kwong, D. Tao, An Underwater Image Enhancement Benchmark Dataset and Beyond, IEEE Transactions on Image Processing, 29 (2020) 4376-4389.
13. A. Singh, T. Hutchinson, X. Wang, Z. Zhang, B. Schafer, F. Derveni, H. Castaneda, K. Peterman, Wall Line Tests: Phase 1 – Shake Table Tests, Designsafe-CI, 2023.