Precise 4D Facial Imaging System Composed by Multiple Infrared Structured Light Sensors and Color Cameras

Di WU[1,†], Yuping YE[2,3,†], Bin REN[1,4], Jixin LIANG[1,4], Zhan SONG[1,5,*]

[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[2]School of Smart Marine Science and Technology, Fujian University of Technology, Fuzhou, China
[3]Fujian Provincial Key Laboratory of Marine Smart Equipment, Fuzhou, China
[4]Southern University of Science and Technology, Shenzhen, China
[5]The Chinese University of Hong Kong, Hong Kong, China

https://doi.org/10.15221/25.26

Abstract

This paper presents a high-precision, multi-view 4D facial imaging system based on structured light. The system comprises three structured-light devices operating at distinct infrared wavelengths. Imaging devices equipped with projectors operating at the 730 nm, 850 nm, and 950 nm bands are deployed at the left-front, front, and right-front positions relative to the face, respectively. Synchronous multi-view scanning is achieved by using distinct wavelength bands combined with an external trigger. In each device, the infrared camera work in conjunction with the projector to reconstruct a 3D point cloud, while the RGB camera at the same view synchronously acquires the corresponding texture. To meet real-time processing requirements, we use CUDA to optimize and accelerate core algorithms, including structured-light decoding, point-cloud meshing, point-cloud reprojection, and normal-map generation. Experiments show that the system maintains a reconstruction accuracy of 0.1 mm while achieving up to 29 Hz full-face 3D reconstruction and high-fidelity rendering exceeding 50 Hz. This provides high-precision, low-latency 3D sensing for real-time applications such as medical diagnosis and virtual reality.

Keywords: structured light, 4D face, multi-view.

1. Introduction

The human face is a primary carrier of biometric identity, affective expression, and physiological function. As one of the most important human biometrics, it is structurally stable, highly distinctive, and amenable to processing, and has therefore received extensive attention across disciplines. Situated at the intersection of computer vision and computer graphics, 3D facial reconstruction aims to recover a face's 3D geometry and texture from 2D images or video sequences. Advances in this technology have provided solutions that overcome the limitations of purely 2D representations in domains such as virtual reality [1,2], biometrics [3,4], and medical diagnosis [5,6]. With the growing demand for dynamic capture, 3D imaging during facial motion (4D facial reconstruction) has emerged as a new technical challenge [7].

However, building a multi-view facial reconstruction system for real-world applications remains challenging. First, non-rigid, rapid, and subtle facial motions induce cross-frame geometric inconsistencies and motion blur. Second, under multi-view settings, active 3D imaging devices suffer fringe-pattern crosstalk. Third, end-to-end real-time performance is difficult to guarantee, as acquisition, transmission, and processing latency directly constrains the speed of 4D facial reconstruction. Finally, there is an inherent trade-off between geometric resolution and rendering detail, making it nontrivial to represent high-frequency features such as wrinkles and pores while meeting real-time requirements. In summary, under the constraints of safety and comfort, reconciling strong real-time performance, high completeness, high-accuracy 3D reconstruction, and high-quality rendering remains a key bottleneck in multi-view 4D facial reconstruction.

According to the principles of data acquisition and the reconstruction methodology, existing 3D facial reconstruction techniques can be broadly divided into 3D imaging methods and deep learning-based methods.

*song.zhan@siat.ac.cn
†Di WU and Yuping YE have contributed equally to this work.

3D imaging approaches are commonly categorized as active or passive. Active scanning either projects specially encoded pattern or emits pulsed illumination, and reconstructs 3D geometry by decoding the projected patterns or analyzing the reflected signals; representative techniques include structured light, speckle-based methods, and time of flight (TOF). Structured light projects encoded patterns onto the object and captures the illuminated regions with a camera; depth is recovered by triangulation. Fringe-structured-light methods [8–10] are efficient and accurate, enabling single-shot surface reconstruction. As a result, many facial 3D imaging systems adopt this method, such as 3dMDface and ABW-3D. Speckle-based 3D imaging systems likewise employ a projectorcamera pair but project a speckle pattern. During reconstruction, the captured speckle is matched to reference patterns and, together with calibration parameters, yields depth. This approach is robust to interference, low-cost, and supports real-time operation [11–13]; representative consumer depth cameras include Microsoft Kinect V1 and the Orbbec Astra series. TOF is a 3D imaging modality based on measuring the travel time of light pulses. The source emits short light pulses toward the target surface; the reflections are captured by the detector, and the time difference between emission and reception yields the distance from the sensor to each surface point. TOF offers a wide field of view, low sensitivity to ambient light, compact form factor, and the ability to sense at longer ranges. Representative consumer depth sensors include Microsoft Kinect V2 and Sony DepthSense 525. Passive 3D scanning requires no active illumination and recovers geometry from multi-view imagery or photometric cues. A representative paradigm is multi-view stereo, which captures images from multiple viewpoints and reconstructs 3D structure from disparities. In practice, many high-precision facial capture platforms are built on this principlefor example, Light Stage series [14] at USC and the Plenoptic Stage at ShanghaiTech University [15]. While these systems deliver accurate geometry and photorealistic textures, their large hardware footprints and costs limit broader deployment.

In recent years, the scale and quality of 3D face datasets [16–18] have increased substantially, spurring rapid progress in deep learning-based 3D facial reconstruction. Among these approaches, those that incorporate parametric face models remain mainstream [19,20]. For example, Wang et al. propose 3DDFA_V3 [21], which segments facial regions and fits the model by optimizing the distribution of selected points, markedly improving a 3DMM's ability to capture extreme expressions. Deng et al. [22] combine pixel-level and semantic-level losses to fuse multi-granularity image cues under weak supervision, enabling single-image 3D face reconstruction by predicting 3DMM coefficients with a CNN. Feng et al. [23] introduce a detail-consistency loss that disentangles fine-scale facial detail from expression, yielding more faithful 3D facial reconstructions on top of FLAME. In addition, Neural Radiance Fields (NeRF) [24–26] and 3D Gaussian Splatting (3DGS) [27–29] have emerged as transformative technologies in 3D reconstruction in recent years, and have been applied to facial reconstruction as well. Gafni et al. [7] propose a learnable dynamic NeRF from monocular video to reconstruct controllable 4D faces. Xu et al. [30] use 3D Gaussians as the primitives of a parametric head model to decouple identity and expression, enabling reconstruction from monocular or few-shot videos. At present, such facial reconstruction methods are beginning to see adoption in film-grade digital-human production. Nevertheless, they still exhibit several limitations, including strong dependence on high-quality facial datasets, limited real-time performance, and insufficient geometric detail fidelity.

In this paper, we present a multi-view, multi-band near-infrared (NIR) structured light 4D facial imaging system. Centered on engineering integration, the system combines cross-device hardware synchronization with GPU parallel acceleration to deliver high-speed 3D reconstruction and high-fidelity rendering of complete faces. We adopt established techniquesstructured-light encoding, triangulation, and normal mappingand introduce targeted optimizations in system architecture, multithreading, and memory management. Experiments show that, while maintaining approximately 0.1 mm geometric accuracy and stable multi-view registration, the system achieves 29 Hz reconstruction and 50 Hz rendering.

Our main contributions are as follows:

- We propose a multi-view, multi-band NIR structured-light 4D facial imaging system. With external triggering and band-pass/IR-cut filtering, it suppresses inter-projector crosstalk and synchronously acquires encoded patterns and RGB textures from three views.
- We design a co-optimized hardware-software parallel pipeline and structured light decoding, point-cloud reprojection, normal map generation, and normal-space transforms on the GPU, enabling real-time 3D reconstruction and high-fidelity rendering.
- We introduce a millisecond-level triangle-mesh generation algorithm based on ordered point clouds and validity masks, improving surface continuity and rendering stability.
- We propose an engineering solution that derives normal maps from high-resolution color images and applies TBN to transform normals from tangent space to world space, enhancing high-frequency surface detail without increasing mesh complexity.

2. Methodology

2.1. Hardware Design

Our system comprises three imaging units positioned at the left front, front, and right front of the face to maximize coverage of the full 3D facial geometry. Each unit operates over a working distance of 0.35 to 0.55 m, with a nominal working distance of 0.45 m within the facial scanning region.
The front unit is pitched slightly downward and reconstructs frontal regions: the mid-forehead, eyes and eyelids, the frontal nasal bridge, frontal cheeks, the mouth and lips, and the midline of the chin. The two lateral units are pitched upward and reconstruct lateral regions, including the temporal region near the temples, the ears, lateral cheeks, the lateral nasal bridge, the lip margins, and the mandible, as shown in Fig. 1b.

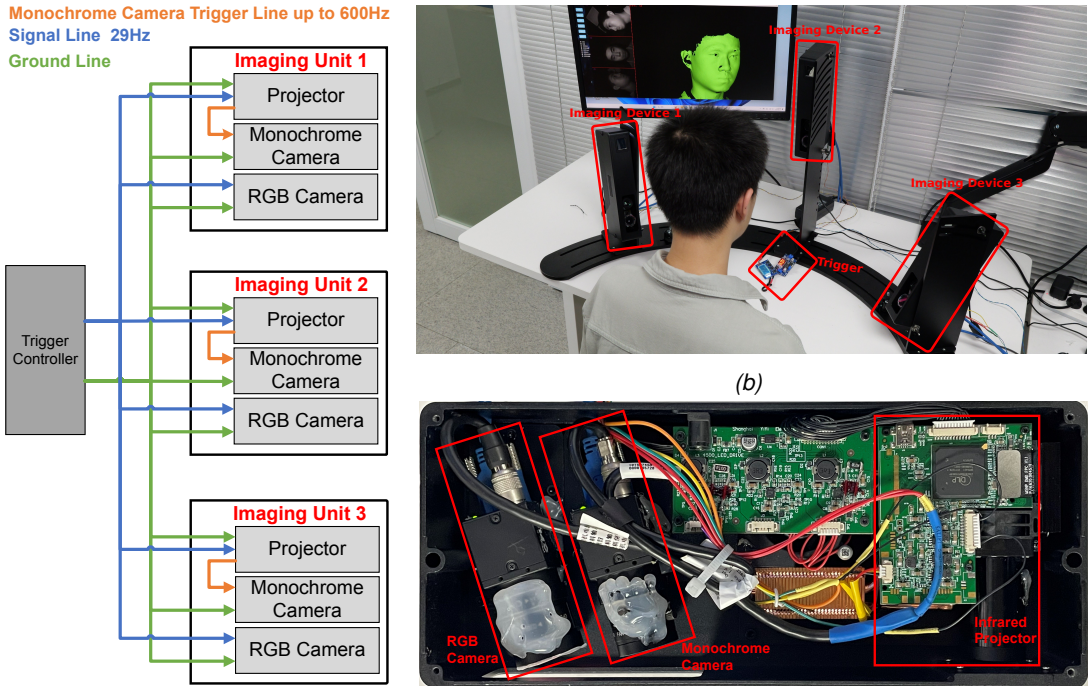


Fig. 1. Device and system configurations: (a) Schematic of external triggering; (b) Complete multi-view real-time structured light system; (c) Single imaging unit configuration

Each imaging unit comprises a Texas Instruments DLP4500 DMD projector (912×1140 pixels; up to 4225 Hz), an industrial camera dedicated to structured-light acquisition (720×540 pixels, 8-bit; up to 600 Hz), and a color camera for RGB texture capture (1440×1080 pixels; up to 76 Hz), as shown in Fig. 1c.
We employ a Gray-code combined with Line-shifting encoding scheme [31]. At a projector horizontal resolution of 1024 pixels, 18 coded patterns are projected. This yields a maximum scanning rate of 29 Hz, given a up to 600 Hz camera rate and 18 patterns.
When multiple devices scan simultaneously, each projector rapidly projects a preset sequence of encoded patterns. To suppress inter-projector crosstalk and to avoid discomfort from high-frequency flicker, we illuminate in three near-infrared bands with center wavelengths of 730 nm, 850 nm, and 950 nm, and place the corresponding band-pass filters in front of the monochrome camera lenses. The sources' spectral power distributions and the filters' transmittance are shown in Fig. 2.
Together with near-infrared narrow-band filters and IR-cut filters, these bands ensure that the monochrome cameras in each subsystem record only the encoded fringes from the in-system projector, while the color cameras do not capture the projected patterns. Although the human eye is sensitive to wavelengths from roughly 380 to 780 nm, its sensitivity around 730 nm is comparatively low. Even high-frequency projection at 730 nm onto the face does not cause noticeable discomfort in practice. An external trigger synchronizes the three devices and the color cameras. It broadcasts a single trigger to all light engines. Each engine projects one preset fringe sequence according to the programmed firmware and, after each projection, outputs a trigger signal to the cameras. The timing is shown in Fig. 1a.

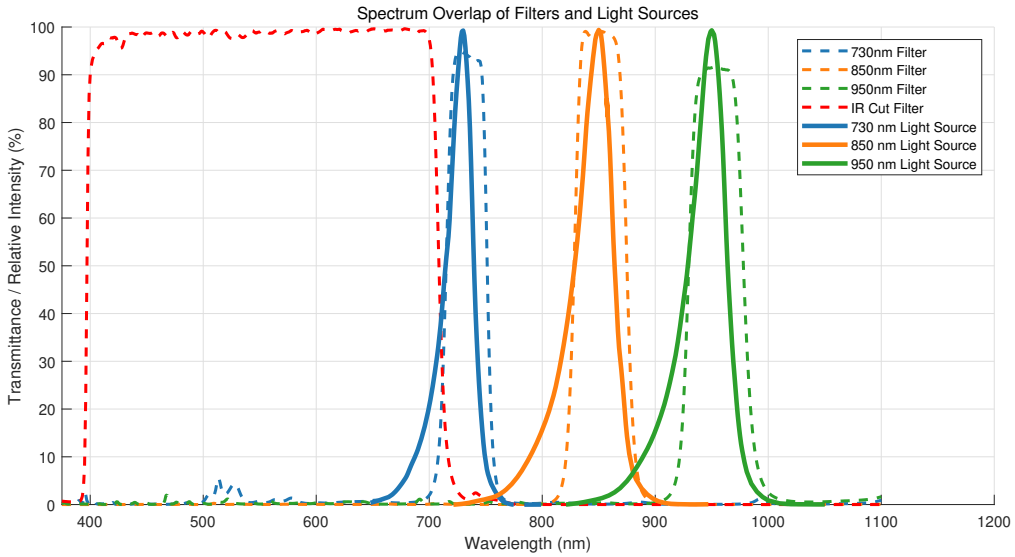


Fig. 2. Schematic diagrams of light source intensity distribution and filter transmittance

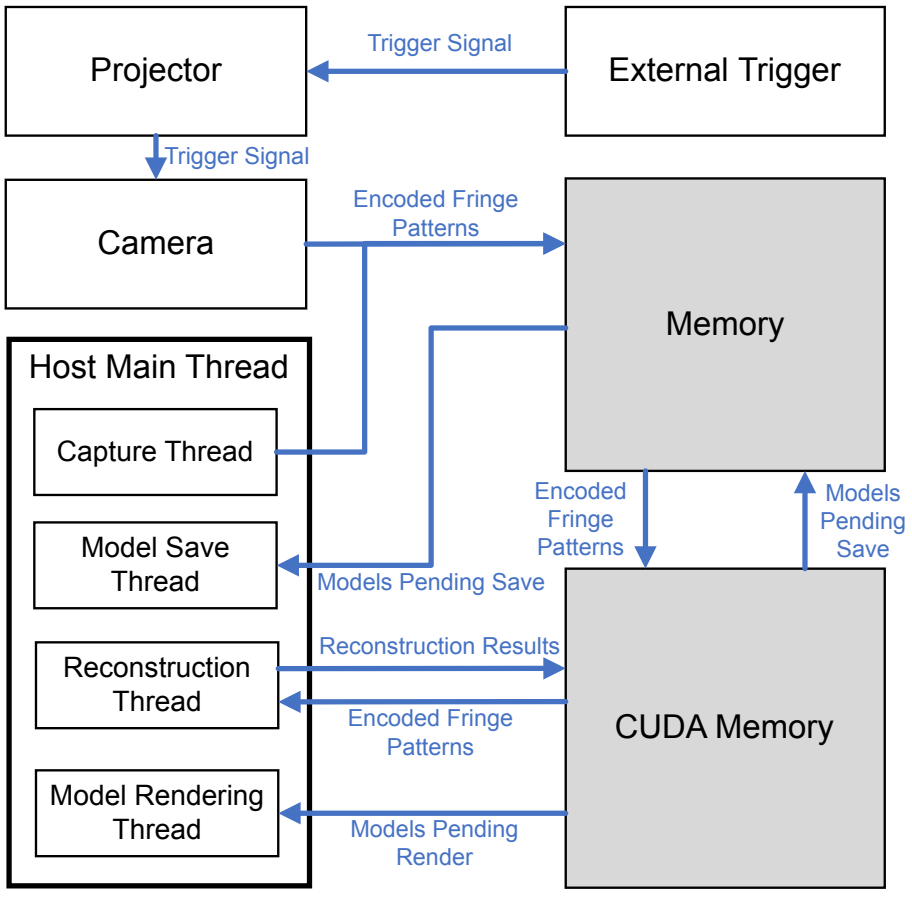


Fig. 3. Software architecture of the real-time, multi-view facial scanning system

2.2. Software Design

Fig. 3 presents the software architecture of our facial scanning system. The pipeline consists of a main host thread and several worker threads: a capture thread, a model-saving thread, a 3D reconstruction thread, and a model-rendering thread.
The capture thread sequentially batches the encoded pattern images acquired from multiple imaging units and writes themvia host memoryinto GPU device memory. The 3D reconstruction thread decodes the captured patterns and, using triangulation, generates the point cloud together with its validity mask. The model-rendering thread performs meshing, computes texture coordinates, and renders the model in real time. The model-saving thread moves models pending persistence from GPU memory to a save queue and commits them to storage in order. Although saving a single model takes much longer than reconstruction, the threads are decoupled; the saving thread does not block the other threads. To operate at the hardware-limited rate (29 Hz full 3D scans per second), we accelerate structured light decoding, point-cloud reprojection, normal map generation, and normal-space transforms using CUDA on the GPU.

2.3. Multi-view Imaging System Calibration

The three structured light devices are rigidly mounted on a common frame, so the relative positions of their point clouds are effectively fixed. We therefore estimate their relative poses by solving the spatial transformation parameters between the point clouds. Because the overlap between views is small, registration based solely on overlapping areas is unreliable for inter-device calibration. We therefore adopt a target-based calibration method.
Specifically, we use a scanner with higher accuracy than our system to acquire a high-precision 3D model of a feature-rich object and create a digital calibration artifact. We then register this digital artifact to the point clouds captured by each subsystem and solve the inter-system transformation matrices. The calibration workflow is shown in Fig.4.

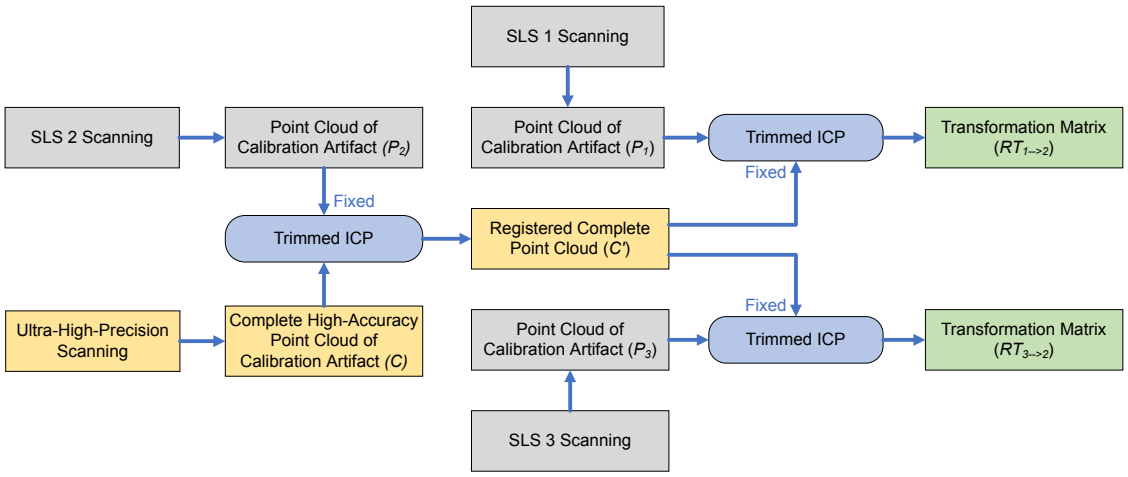


Fig. 4. Calibration workflow of the multi-view system based on a calibration artifact

2.4. Model Rendering

To evaluate reconstruction quality in real time, we render the streaming facial model as it is acquired. Common 3D representations include point clouds, triangle meshes, and quad meshes, which differ markedly in data structures, typical applications, and processing. Point clouds are sets of samples without explicit surfaces and therefore appear grainy compared with meshes. For a fixed number of vertices, quad meshes are less expressive geometrically than triangle meshes. Accordingly, we render models as triangle meshes and employ normal mapping together with simultaneously captured textures to achieve real-time, high-fidelity facial rendering.

2.4.1. Triangulation Algorithms for Ordered Point Clouds

An ordered point cloud arranges samples on a regular grid by rows and columns. Unlike unorganized point clouds that require k-d trees to query neighborhoods, neighbors can be accessed directly by rowcolumn indices.

In addition, the mask produced by fringe-structured light marks valid regions, where each pixel with a true mask corresponds to a valid 3D point. We leverage the ordered point cloud and its mask to generate triangle meshes efficiently.
The core idea is to traverse every 2×2 pixel block in the mask and determine the triangles according to the valid-point pattern. As shown in Fig.5, when at least three valid points exist in a block, we form different triangle meshes depending on the location of the invalid point. For correct rendering, triangles follow a counterclockwise winding order, as specified below:

- All four points are valid: generate $\triangle ACB$ and $\triangle CDB$
- Point A is invalid: generate $\triangle CDB$
- Point B is invalid: generate $\triangle ACD$
- Point C is invalid: generate $\triangle ADB$
- Point D is invalid: generate $\triangle ACB$

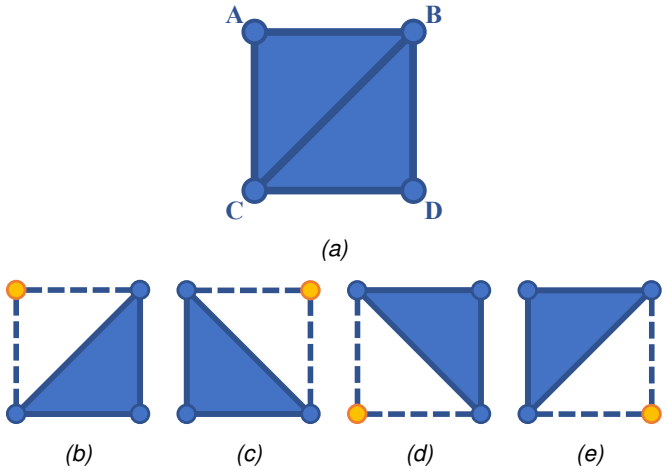


Fig. 5. Triangles to be constructed in the pixel grid (blue dots indicate valid points, yellow dots indicate invalid points): (a) All points are valid; (b) Point A is invalid; (c) Point B is invalid; (d) Point C is invalid; (e) Point D is invalid

We set a distance threshold to prevent triangles from connecting spatially disconnected regions that may appear contiguous in the mask. If any edge of a candidate triangle exceeds this threshold, the triangle is discarded. The algorithm's time complexity depends only on the monochrome camera resolution and is $\mathcal{O}(w \times h)$. Consequently, a single pass over all pixel blocks in the mask suffices to generate the mesh.

2.4.2. Normal Map Generation

Normal mapping simulates fine surface relief by modulating lighting. During rendering, a low-resolution mesh samples a normal map baked from a high-resolution model; the fragment shader fetches modified normals from the texture using UV coordinates for lighting computations. This technique enhances surface detail without changing the vertex count or mesh topology by altering per-fragment surface normals.
In this work, because no higher-precision model is available, we enhance the rendering of the structured light reconstruction by generating normal maps from high-resolution 2D images captured by the color camera (1440×1080, higher than the monochrome camera's 720×540), rather than by a conventional high-to-low baking pipeline. The steps are as follows:

1) **Height map generation**. We assume a positive correlation between pixel intensity $I_{x,y}$ and surface height $H_{x,y}$, so the color image is first converted to grayscale using a perception-based transform, yielding the height map:

$$I_{x,y} = 0.2989 \cdot Color^{r}_{x,y} + 0.5870 \cdot Color^{g}_{x,y} + 0.1140 \cdot Color^{b}_{x,y} \tag{1}$$

$$H_{x,y} = amplitude \cdot \frac{I_{x,y} - min(I)}{max(I) - min(I)} \tag{2}$$

Here, $amplitude$ scales the height variation; larger values produce stronger relief contrast. Coordinates x and y denote the pixel location in the image.

2) **Multi-directional sampling coordinates**. For each pixel, we define eight principal directions with angles $a \cdot \frac{\pi}{4}$. For each direction, we pick three samples with coordinates $(sx^a_{1,2,3}, sy^a_{1,2,3})$:

$$\theta^a_{base} = a \cdot \frac{\pi}{4} \tag{3}$$

$$(sx^a_1, sy^a_1) = (\text{int}\,(x + r \cdot \cos\theta^a_{base})\,, \text{int}\,(y + r \cdot \sin\theta^a_{base})) \tag{4}$$

$$(sx^a_2, sy^a_2) = \left(\text{int}\left(x + r \cdot \cos\left(\theta^a_{base} + \frac{\pi}{3}\right)\right), \text{int}\left(y + r \cdot \sin\left(\theta^a_{base} + \frac{\pi}{3}\right)\right)\right) \tag{5}$$

$$(sx^a_3, sy^a_3) = \left(\text{int}\left(x + r \cdot \cos\left(\theta^a_{base} + \frac{2\pi}{3}\right)\right), \text{int}\left(y + r \cdot \sin\left(\theta^a_{base} + \frac{2\pi}{3}\right)\right)\right) \tag{6}$$

The parameter r is the sampling radius. Increasing r captures larger-scale relief and yields smoother normals across neighboring pixels.

3) **Normal computation**. For each principal direction, compute $h^a_1 \times h^a_2$ from vectors linking the first sample to the other two. Sum all directional normals and normalize to obtain the pixel's normal $n_{p,q}$:

$$h^a_1 = (sx^a_2 - sx^a_1, sy^a_2 - sy^a_1, H_{sx^a_2, sy^a_2} - H_{sx^a_1, sy^a_1}) \tag{7}$$

$$h^a_2 = (sx^a_3 - sx^a_1, sy^a_3 - sy^a_1, H_{sx^a_3, sy^a_3} - H_{sx^a_1, sy^a_1}) \tag{8}$$

$$n_{p,q} = \text{normalize} \sum_{a=0}^{7} (h^a_1 \times h^a_2) \tag{9}$$

Increasing the sampling radius r produces smoother normal maps, whereas increasing $amplitude$ yields more pronounced relief. In addition, because the RGB channels encode the (X, Y, Z) components of normals and most normals point upward in the image, these maps appear predominantly blue.

2.4.3. Tangent Space Transformation

Tangent space, like model space and world space, is a coordinate frame used to efficiently process surface detail and lighting. It comprises three orthonormal unit vectors: the tangent, bitangent, and normal. To keep texture maps reusable and invariant under global transforms, the normal offsets encoded in a normal map are represented in tangent space rather than in model or world space.
In practice, normals stored in the normal map are first transformed from tangent space to the world space of the model, then lighting is computed. As with other coordinate transforms, this uses the TBN matrix (tangent, bitangent, normal). For any triangle with vertices $P_1(U_1, V_1)$, $P_2(U_2, V_2)$, and $P_3(U_3, V_3)$, the tangent and bitangent of the tangent frame can be derived from the UVs. In Fig. 6, the red and green arrows denote the tangent and bitangent, and the normal is perpendicular to the triangle plane.

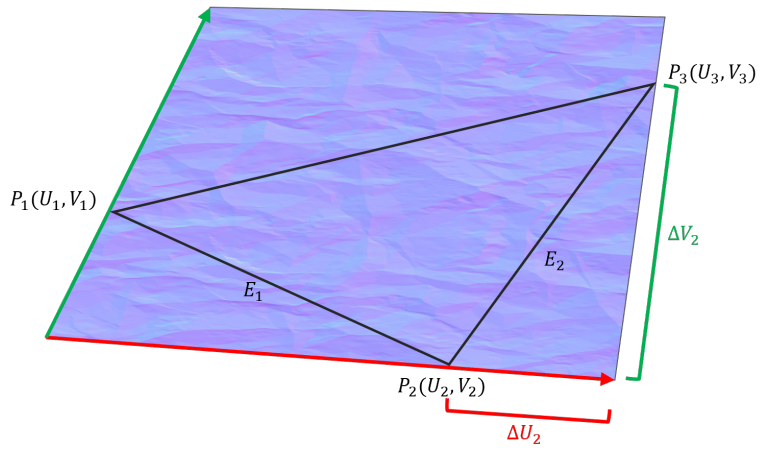


Fig. 6. Definition of tangent space (Red: tangent vector; Green: bitangent vector)

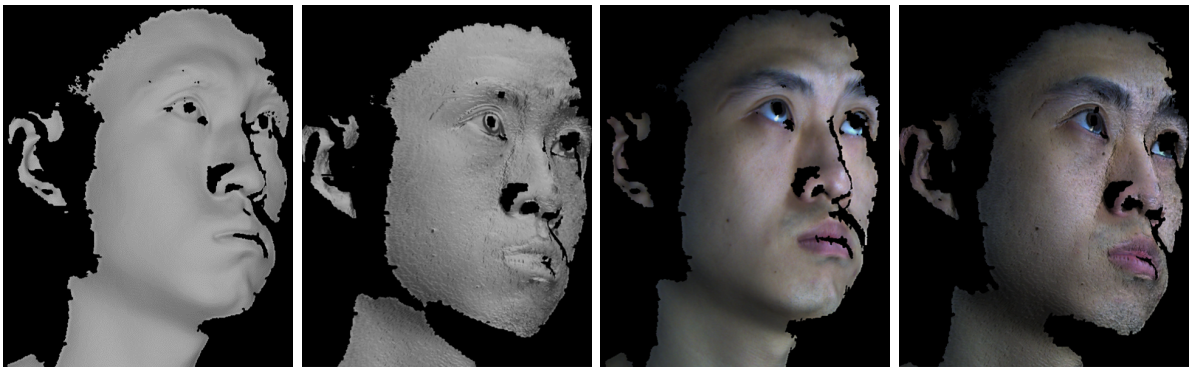
First, express the triangle edge vectors E_1 and E_2 as linear combinations of the tangent T and bitangent B:

$$\begin{bmatrix} E_{1x} & E_{1y} & E_{1z} \\ E_{2x} & E_{2y} & E_{2z} \end{bmatrix} = \begin{bmatrix} \Delta U_1 & \Delta V_1 \\ \Delta U_2 & \Delta V_2 \end{bmatrix} \begin{bmatrix} T_x & T_y & T_z \\ B_x & B_y & B_z \end{bmatrix} \tag{10}$$

Then, using E_1, E_2, and the texture coordinates, compute the triangle's tangent T and bitangent B, which are shared by all points on the triangle:

$$\begin{bmatrix} T_x & T_y & T_z \\ B_x & B_y & B_z \end{bmatrix} = \frac{1}{\Delta U_1 \Delta V_2 - \Delta U_2 \Delta V_1} \begin{bmatrix} \Delta V_2 & -\Delta V_1 \\ -\Delta U_2 & \Delta U_1 \end{bmatrix} \begin{bmatrix} E_{1x} & E_{1y} & E_{1z} \\ E_{2x} & E_{2y} & E_{2z} \end{bmatrix} \tag{11}$$

Typically, a vertex belongs to multiple triangles. For smoother results, per-vertex tangents and bitangents are averaged across adjacent triangles. After computing per-vertex T and B, the TBN matrix is constructed in the vertex shader and passed to the fragment shader. Finally, normals from the normal map (given in tangent space) are transformed into world space via TBN and used with lighting quantities defined in the same space. As an example with a scanned face, Fig.7 compares renderings with and without texture and normal maps. Texture alone increases realism, but lacks relief. With a normal map, high-resolution texture provides additional detail through normal perturbation.



(a) (b) (c) (d)

Fig. 7. Comparative renderings under different texture and normal mapping conditions: (a) no texture or normal map; (b) normal map only; (c) texture only; (d) texture plus normal map

3. Experimental Results

Our experiments were conducted on a workstation with an Intel(R) Core(TM) i9-14900KF CPU, an NVIDIA GeForce RTX 4080 SUPER GPU with 16 GB VRAM, and 64 GB system memory.

3.1. System Performance

We profiled per-capture runtimes on the GPU for structured light decoding, point-cloud reprojection, normal map generation, and normal-space transforms, as summarized in Table 1.

Table 1. Execution time of operations with CUDA acceleration

Processing step	Grid size	Block size	Time (ms)
Structured light decoding	23×17	32×32	1.4426
Point-cloud reprojection	45×34	16×16	0.2858
Normal map generation	60×34	32×32	0.9022
Normal-space transform	60×34	32×32	0.0084

In practice, CUDA parallelization combined with multithreading raises the 3D reconstruction rate to 29 Hz, which matches the monochrome camera's hardware limit. With normal mapping disabled, the rendering frame rate remains above 70 Hz; with normal mapping enabled, it still exceeds 50 Hz. Fig. 8 shows three sequences of real-time rendered 3D facial models.

3.2. Imaging Precision Verification

We validate accuracy using a precision planar target with 5 μm manufacturing tolerance and equidistant markers, as shown in Fig.9a. Red arrows indicate the distance and angular metrics used for evaluation. Fig.9b, 9c and 9d visualize the spatial distribution of plane-fitting errors. Table 2 reports quantitative results for

Fig. 8. Real-time rendered 3D facial model

Table 2. Planar geometric accuracy of three imaging units

Criteria	Min	Max	Mean	Std. dev
	730 / 850 / 950	**730 / 850 / 950**	**730 / 850 / 950**	**730 / 850 / 950**
Distance (mm)	-0.131 / -0.152 / -0.090	0.354 / 0.232 / 0.406	0.016 / 0.011 / 0.026	0.025 / 0.030 / 0.036
Angle (°)	-0.070 / -0.103 / -0.101	0.169 / 0.131 / 0.123	0.038 / 0.058 / 0.042	0.087 / 0.067 / 0.073
Planarity (mm)	-0.236 / -0.190 / -0.193	0.201 / 0.161 / 0.137	-0.011 / 0.029 / 0.018	0.070 / 0.057 / 0.042

distance accuracy, angular accuracy, and planarity. The experiments show that all three real-time structured light systems built in this chapter reconstruct objects with high accuracy.

3.3. Multi-view Imaging System Calibration Results

Following the target-based multi-view calibration in section 2.3, we used a GCI PARDUS P400 3D imaging device to scan a plaster bust from multiple viewpoints and registered the scans into a complete model. This device is also based on structured light technique, offering 3.2M resolution and 8 μm repeatability along the Z axis, both exceeding those of our system. The calibration artifact and its high-precision reference model are shown in Fig.10. We then placed the bust at the center of the enclosure and captured a multi-view dataset with our system, after which we registered the three views to estimate their poses. The per-view scans and the error distributions relative to the reference are shown in Fig.11b,11c and 11d. After system calibration, the stitched point-cloud results from all views are shown in Fig. 11a.

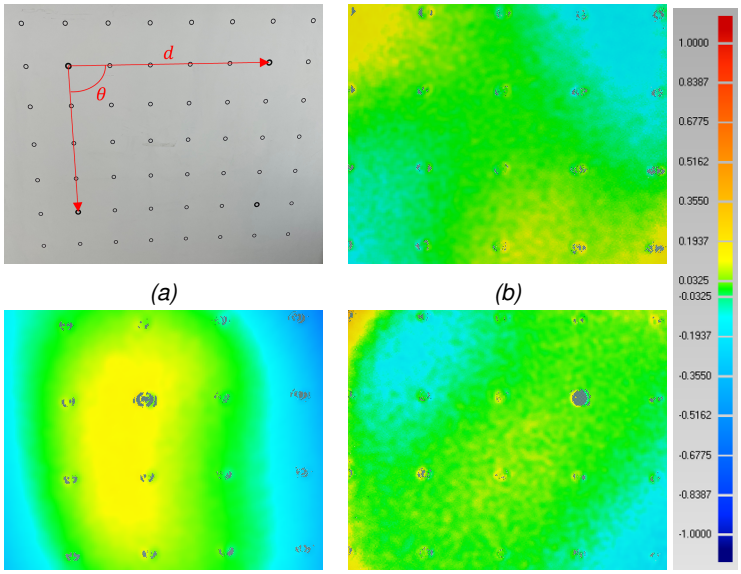


Fig. 9. Planarity distribution of multi-view structured light imaging system: (a) High-precision calibration board; (b) Imaging unit 1 (730 nm); (c) Imaging unit 2 (850 nm); (d) Imaging unit 3 (950 nm)

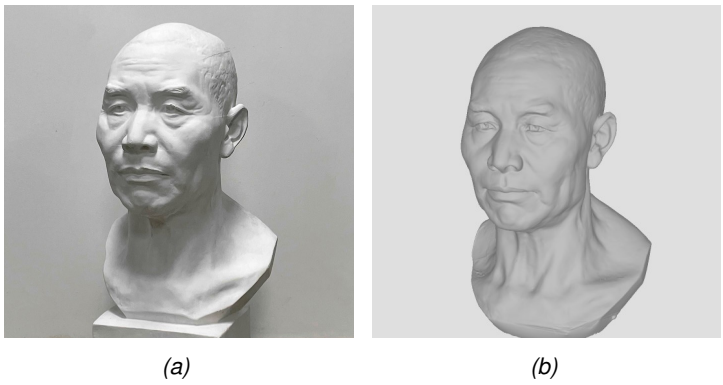


Fig. 10. System calibration objects: (a) Plaster statue; (b) Corresponding 3D model

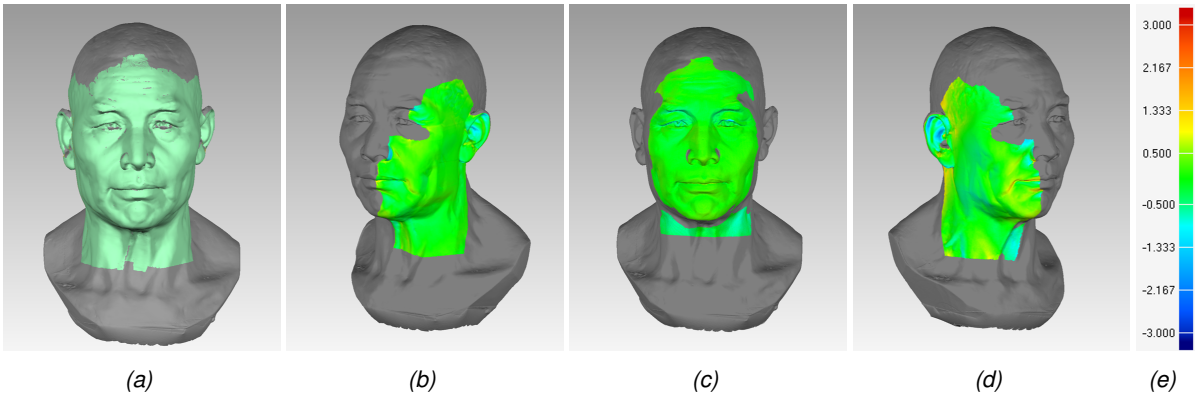


Fig. 11. Multi-view registration result and per-view error distributions with respect to the high-precision reference model: (a) merged point clouds from three views; (b) error distribution of the 730 nm unit; (c) error distribution of the 850 nm unit; (d) error distribution of the 950 nm unit; (e) color scale of error magnitude

4. Conclusions

We present a multi-view, multi-band structured light 4D facial imaging system. On the hardware side, external triggering together with narrow band-pass and IR-cut filtering suppress inter-device crosstalk and enable synchronized acquisition of encoded patterns and RGB textures from three views. On the software side, a worker-thread parallel pipeline with CUDA acceleration across multiple stages significantly reduces end-to-end latency. On a workstation with an Intel i9-14900KF CPU and an NVIDIA RTX 4080 SUPER GPU, the system achieves 29 Hz full 3D reconstruction and over 50 Hz high-fidelity rendering (over 70 Hz with enhancement disabled), while delivering approximately 0.1 mm geometric accuracy and stable multi-view registration in planar and bust calibrations. In the future, the system can further reduce cost and latency and improve robustness and usability, enabling higher-quality 4D facial sensing for medical diagnosis, rehabilitation assessment, human-computer interaction, and digital-human production

This work is supported by the Shenzhen High-tech Zone Development Special Plan Innovation Platform Construction.

References

[1] P. Yan, R. K. Ward, Q. Tang, and S. Du, “Neural 3d face shape stylization based on single style template via weakly supervised learning,” *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[2] S.-Y. Chen, Y.-K. Lai, S. Xia, P. L. Rosin, and L. Gao, “3d face reconstruction and gaze tracking in the hmd for virtual interaction,” *IEEE Transactions on Multimedia*, vol. 25, pp. 3166–3179, 2022.

[3] F. Liu, Q. Zhao, X. Liu, and D. Zeng, “Joint face alignment and 3d face reconstruction with application to face recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 664–678, 2018.

[4] M. He, J. Zhang, S. Shan, and X. Chen, “Enhancing face recognition with self-supervised 3d reconstruction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4062–4071.

[5] W. Kong, Z. You, S. Lyu, and X. Lv, “Multi-dimensional stereo face reconstruction for psychological assistant diagnosis in medical meta-universe,” *Information Sciences*, vol. 654, p. 119831, 2024.

[6] M. A. Alagha, A. Ayoub, S. Morley, and X. Ju, “Objective grading facial paralysis severity using a dynamic 3d stereo photogrammetry imaging system,” *Optics and Lasers in Engineering*, vol. 150, p. 106876, 2022.

[7] G. Gafni, J. Thies, M. Zollhofer, and M. Nießner, “Dynamic neural radiance fields for monocular 4d facial avatar reconstruction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8649–8658.

[8] X. Meng, P. Zhang, S. Wang, and B. Lu, “Three-dimensional reconstruction algorithm for facial plastic surgery using high-precision monocular structured light,” in *2022 16th ICME International Conference on Complex Medical Engineering (CME)*. IEEE, 2022, pp. 341–344.

[9] Z. Wang, “Robust three-dimensional face reconstruction by one-shot structured light line pattern,” *Optics and Lasers in Engineering*, vol. 124, p. 105798, 2020.

[10] F. Bruno, G. Bianco, M. Muzzupappa, S. Barone, and A. V. Razionale, “Experimentation of structured light and stereo vision for underwater 3d reconstruction,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 4, pp. 508–518, 2011.

[11] Y. Li, J. Shi, S. Wu, and D. Ye, “High-fidelity 3d reconstruction via global gradient-optimized speckle patterns and color-guided stereo matching with subpixel error suppression,” *Measurement*, p. 118250, 2025.

[12] K. Fu, Y. Xie, H. Jing, and J. Zhu, “Fast spatial–temporal stereo matching for 3d face reconstruction under speckle pattern projection,” *Image and Vision Computing*, vol. 85, pp. 36–45, 2019.

[13] P. Zhou, J. Zhu, and H. Jing, “Optical 3-d surface reconstruction with color binary speckle pattern encoding,” *Optics express*, vol. 26, no. 3, pp. 3452–3465, 2018.

[14] K. Guo, P. Lincoln, P. Davidson, J. Busch, X. Yu, M. Whalen, G. Harvey, S. Orts-Escolano, R. Pandey, J. Dourgarian *et al.*, "The relightables: Volumetric performance capture of humans with realistic relighting," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 6, pp. 1–19, 2019.

[15] L. Zhang, Q. Zhang, M. Wu, J. Yu, and L. Xu, "Neural video portrait relighting in real-time via consistency modeling," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 802–812.

[16] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 601–610.

[17] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3d face shape and expression from an image without 3d supervision," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7763–7772.

[18] Y. Ye, Z. Song, J. Guo, and Y. Qiao, "Siat-3dfe: a high-resolution 3d facial expression dataset," *IEEE Access*, vol. 8, pp. 48 205–48 211, 2020.

[19] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187–194.

[20] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans." *ACM Trans. Graph.*, vol. 36, no. 6, pp. 194–1, 2017.

[21] Z. Wang, X. Zhu, T. Zhang, B. Wang, and Z. Lei, "3d face reconstruction with the geometric guidance of facial part segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1672–1682.

[22] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[23] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–13, 2021.

[24] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 318–10 327.

[25] A. Chen, Z. Xu, F. Zhao, X. Zhang, F. Xiang, J. Yu, and H. Su, "Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 124–14 133.

[26] Y. Hong, B. Peng, H. Xiao, L. Liu, and J. Zhang, "Headnerf: A real-time nerf-based parametric head model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 374–20 384.

[27] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[28] D. Qin, H. Lin, Q. Zhang, K. Qiao, L. Zhang, J. Saito, Z. Zhao, J. Yu, L. Xu, and T. Komura, "Instant gaussian splatting generation for high-quality and real-time facial asset rendering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[29] Z. Qian, S. Wang, M. Mihajlovic, A. Geiger, and S. Tang, "3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 5020–5030.

[30] Y. Xu, Z. Su, Q. Wu, and Y. Liu, "Gphm: Gaussian parametric head model for monocular head avatar reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[31] Y. Ye, S. Hao, Z. Song, F. Gu, and J. Zhao, "A novel triangular stereo model for 3d reconstruction of uniaxial mems-based structured light system," *Optics and Lasers in Engineering*, vol. 166, p. 107596, 2023.