

Article

A High-Fidelity Facial Digital Twin Benchmark for Quantitative Evaluation of AI-Based 3D Eye Tracking

Kaiqiao Tian ^{1,*}, Mohammad S. Alzyout ¹, Zhengyi Lu ¹, Changqing Cai ^{2,*}, Khalid Mirza ¹, Ka C. Cheok ¹ and Shadi Alawneh ¹

¹ Electrical and Computer Engineering, Oakland University, Rochester, MI 48309, USA; malzyout@oakland.edu (M.S.A.); zhengyilu@oakland.edu (Z.L.); mirza@oakland.edu (K.M.); cheok@oakland.edu (K.C.C.); shadi.alawneh@oakland.edu (S.A.)

² College of Electrical and Information Engineering, Changchun Institute of Technology, Changchun 130012, China

* Correspondence: tian2@oakland.edu (K.T.); caicq@ccit.edu.cn (C.C.)

Abstract

Artificial intelligence (AI)-based 3D eye tracking is a fundamental enabling technology for human–computer interaction (HCI), extended reality (XR), and intelligent spatial computing. However, physical evaluation methodologies are heavily limited by non-repeatable human micro-movements, the lack of precise millimeter-level ground truth, and systemic camera calibration errors. Addressing these limitations, this paper delivers a methodological meta-contribution to the field by presenting a high-fidelity facial digital twin benchmark framework using NVIDIA Isaac Sim for the rigorous and repeatable evaluation of 3D eye-tracking algorithms. Rather than focusing on incremental algorithmic modifications, our framework establishes a standardized, hardware-free testing paradigm. By systematically sampling virtual facial poses under identical rendering configurations, it generates dense, fully repeatable trajectories with mathematically exact spatial ground truth across landmark-based, 3DMM-based, and direct regression architectures. To isolate intrinsic algorithmic capabilities from extrinsic calibration biases, we propose a self-referenced relative-motion evaluation protocol operating in a facial-centered local reference frame. Comprehensive diagnostics are performed across multiple key dimensions, including localization accuracy, temporal jitter, detection robustness, and pose sensitivity, culminating in a newly introduced Comprehensive Performance Index (CPI) to aggregate these multi-dimensional metrics. Statistical hypothesis testing reveals consistent performance hierarchies: landmark and 3DMM methods achieve superior geometric consistency and temporal stability by leveraging parametric shape constraints, whereas direct regression models exhibit severe tracking degradation and failure under extreme rotations. By resolving the long-standing benchmark replication bottleneck, this extensible digital twin platform establishes a standardized, reproducible methodology for developing and certifying trustworthy human-centric perception systems.



Academic Editors: Panayiotis Bozanis and Leonidas Akritidis

Received: 31 May 2026

Revised: 13 July 2026

Accepted: 22 July 2026

Published: 25 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: artificial intelligence (AI); 3D eye tracking; facial digital twin; extended reality (XR); virtual reality (VR); augmented reality (AR); quantitative evaluation; relative-motion evaluation; reliability assessment; human–computer interaction (HCI); NVIDIA Isaac Sim; human-centric perception

1. Introduction

1.1. Background and Motivation

Next-generation interactive technologies, including extended reality (XR), spatial computing, intelligent vehicles, and human–robot collaboration, increasingly rely on robust AI-based human-centric perception to enable natural, seamless, and adaptive human–computer interaction (HCI) [1–3]. Among these perception capabilities, 3D eye and gaze tracking has become a fundamental component for numerous applications, including gaze-contingent rendering, hands-free interaction, user intent prediction, digital avatar animation, driver monitoring, and collaborative robotics [4]. As AI-driven perception systems continue to expand into safety-critical domains, evaluating the operational reliability of these algorithms has become just as critical as enhancing their raw tracking accuracy.

Recent advances in deep learning have significantly improved the performance of 3D eye-tracking algorithms. Representative approaches span several paradigms, including classical feature-based landmark detectors [5], dense 3D morphable model (3DMM)-based geometric regressors [6,7], and direct regression architectures using facial bounding boxes or raw pixel coordinates [8–10]. Although these methods achieve impressive accuracy on standardized training datasets, their robustness under out-of-plane head poses, extreme viewing angles, and varying sensor conditions remain insufficiently understood. In many practical deployment scenarios, unexpected tracking failures, temporal instability, and geometric inconsistencies can severely degrade the user experience, interaction quality, and overall system safety.

Despite rapid algorithmic progress, the evaluation methodologies used to assess these systems have evolved much more slowly. Traditionally, landmark-based and appearance-based eye-tracking models are validated on large-scale public datasets such as Gaze-Capture [11] and ETH-XGaze [12]. While these datasets are invaluable for training and general performance estimation, they are primarily static collections that rely on absolute-coordinate comparisons. Consequently, the reported error metrics are heavily contaminated by extrinsic camera calibration inaccuracies, sensor mounting tolerances, and coordinate alignment offsets. Furthermore, evaluations performed on human participants are inherently limited by natural, non-repeatable micro-movements, making it virtually impossible to establish a clean, hardware-independent baseline [13]. These confounding factors make it extremely difficult to isolate the intrinsic capabilities of competing algorithms.

To address these limitations, there is an urgent need for a standardized, reproducible, and mathematically rigorous evaluation framework. It is critical to recognize that in a field heavily bottlenecked by non-standardized benchmarks and irreproducible physical experimental setups, establishing a rigorous evaluation methodology constitutes a fundamental “meta-contribution” to the scientific community. The core objective of this work is not to propose yet another incremental 3D tracking algorithm, but rather to construct and validate this very methodological paradigm. By providing a virtual surrogate that isolates intrinsic algorithmic vulnerabilities from physical sensing dependencies and calibration biases, this framework delivers a standardized, reproducible foundation essential for stress-testing, diagnosing, and certifying trustworthy human-centric perception systems.

1.2. Challenges in Reliability Evaluation

Establishing a rigorous methodology to evaluate AI-based 3D eye-tracking algorithms faces several fundamental challenges. In real-world experimental setups, multiple sources of physical and geometric uncertainty obscure the actual performance of the algorithms under test. The primary challenges can be summarized into three main areas:

First, there is a fundamental lack of repeatable motion and precise ground truth. Human head and eye movements are inherently variable. A human subject cannot replicate a

spatial trajectory with millimeter-level precision across repeated trials. Moreover, obtaining precise 3D ground truth of the eye centers in physical space is highly intrusive. Standard motion-capture systems require physical markers or specialized hardware mounted on the subject, which alters the natural facial appearance and directly interferes with the AI model's keypoint detection.

Second, absolute coordinate evaluations are highly sensitive to systemic calibration bias. Conventional evaluation protocols compare model predictions against ground truth in an absolute world coordinate frame. This approach makes the resulting error metrics extremely sensitive to camera intrinsic calibration errors, hand–eye alignment offsets, and coordinate system discrepancies. These systematic biases often exceed the intrinsic tracking errors of the models themselves, thereby invalidating comparative benchmarking across different platforms.

Third, hardware-induced sensor noise complicates error isolation. Active depth sensors, such as RGB-D or time-of-flight (ToF) cameras, introduced depth quantization noise, multipath interference, and edge-artifacts [14]. These physical sensing limitations propagate directly into the 3D projection pipeline, generating high-frequency spatial jitter. Without a controlled environment to isolate these variables, it is impossible to determine whether tracking degradation is caused by algorithmic failure or physical sensor limitations.

1.3. High-Fidelity Facial Digital Twin Benchmark

High-fidelity simulation platforms, such as NVIDIA Isaac Sim, provide a powerful solution to these physical benchmarking limitations. Rather than serving merely as a tool for synthetic data generation, a facial digital twin within a virtual environment function as a standardized, hardware-independent evaluation surrogate. This approach allows researchers to independently control and vary environmental parameters without introducing physical calibration errors.

Recent literature demonstrates the growing value of synthetic face analysis and digital prototyping for computer vision models [15–17]. In particular, recent frameworks like UnityEyes [18,19], UnityEyes 2 [20], and digital hardware prototyping models [21] have shown that high-fidelity simulations can accurately model camera optics, sensor noise, and complex spatial geometries. By leveraging these platforms, a virtual human face can be systematically sampled across dense translational and rotational spaces while providing mathematically exact spatial ground truth.

For benchmarking, this simulation paradigm offers several unique advantages. First, it enables absolute repeatability; identical facial poses and trajectories can be executed indefinitely under identical illumination and camera configurations. Second, it allows for extreme case testing. Challenging boundary conditions, such as extreme out-of-plane head rotations and severe self-occlusions, can be safely and consistently evaluated, which is practically impossible with human participants. This capability is vital for identifying the precise failure boundaries of AI-based perception models.

1.4. Research Gap

Despite the clear benefits of simulation-based validation, several critical gaps remain in the current eye-tracking literature:

1. Lack of standardized, repeatable evaluation benchmarks: While significant effort has been dedicated to synthesizing training data to improve model accuracy, there is a notable absence of standardized virtual benchmarks designed specifically for repeatable, multi-axis reliability diagnostics. The field lacks a formalized evaluation methodology to bridge simulation environments and algorithm verification.

2. Calibration-induced evaluation bias: Existing evaluation protocols rely almost exclusively on absolute coordinate systems, where extrinsic alignment errors mask the true, intrinsic tracking capabilities of competing models.
3. Lack of deep cross-architecture benchmarking: There is no systematic comparison evaluating how different architectural paradigms—such as classical geometry, lightweight convolutional detectors, and heavy 3DMM regressors—behave under identical, highly controlled motion profiles.
4. Insufficient analysis of boundary conditions: Most benchmarks report only global average errors, leaving critical performance characteristics like temporal jitter, depth-dependent degradation, and rotational failure boundaries largely unquantified.

To the best of our knowledge, no prior work has established a high-fidelity facial digital twin benchmark that combines dense multi-axis pose sampling, mathematically exact ground truth, and a calibration-free evaluation protocol to systematically diagnose and compare different AI-based 3D eye-tracking architectures.

1.5. Research Questions

To address these research gaps, this study is designed to answer the following research questions:

RQ1: Can a high-fidelity facial digital twin provide a standardized, hardware-independent, and fully repeatable evaluation methodology for diagnosing AI-based 3D eye-tracking algorithms?

RQ2: Can a self-referenced relative-motion evaluation protocol mathematically eliminate camera calibration and hand–eye alignment biases, thereby isolating the intrinsic tracking capabilities of the models?

RQ3: How do different architectural paradigms—specifically landmark-based, 3DMM-based, and bounding-box-based methods—compare in terms of localization accuracy, temporal stability, and rotational robustness under identical evaluation conditions?

RQ4: What are the characteristic failure modes, operational boundaries, and multi-dimensional rankings of these representative architectures when subjected to dense translational and rotational pose variations?

1.6. Main Contributions

The primary contributions of this work are summarized as follows:

1. A Standardized Benchmarking Methodology Based on a High-Fidelity Facial Digital Twin: Rather than proposing an incremental tracking algorithm, we establish a standardized, highly reproducible evaluation methodology for 3D human-centric perception. By leveraging a high-fidelity facial digital twin in NVIDIA Isaac Sim, this framework provides dense, multi-axis pose sampling, mathematically exact spatial ground truth, and absolute experimental repeatability, resolving the long-standing benchmark replication bottleneck in physical setups.
2. A Calibration-Free, Self-Referenced Relative Evaluation Protocol: We propose a relative-motion evaluation protocol that shifts the frame of reference from absolute world coordinates to a local, facial-centered coordinate system. Unlike conventional absolute evaluations, this formulation mathematically eliminates camera calibration and hand–eye alignment offsets, ensuring a fair and isolated assessment of model performance.
3. A Unified Multidimensional Evaluation Framework and Comprehensive Index (CPI): We introduce a unified evaluation framework that systematically measures localization accuracy, temporal stability, robustness, and poses sensitivity under identical

controlled conditions, enabling fair and reproducible comparison across different eye-tracking architectures.

4. Deep Cross-Architecture Diagnostics and Failure Discovery: We conduct a systematic evaluation of five representative eye-tracking paradigms under identical virtual conditions. Our analysis exposes distinct, architecture-specific failure modes, including the catastrophic rotational tracking collapse of global bounding-box regressors, providing clear design guidelines for future robust perception systems.

Importantly, the primary contribution of this work is not the proposal of a new eye-tracking algorithm, but the establishment of a standardized and reproducible evaluation methodology for the field. In the absence of reliable benchmark protocols, methodological frameworks themselves constitute an essential scientific contribution, because they define how algorithmic performance should be measured, compared, and interpreted.

2. Facial Digital Twin and Simulation Framework

To establish a highly controllable and mathematically rigorous evaluation platform, we constructed a high-fidelity facial digital twin and virtual sensor testing framework within NVIDIA Isaac Sim. This virtual testbed eliminates physical hand–eye calibration errors and human-motion variability, enabling the direct isolation of intrinsic algorithmic vulnerabilities.

High-fidelity reconstruction of human facial geometry has been an active research area driven by applications in visual effects, biometric authentication, medical imaging, and digital human modeling [22–24]. Ghosh et al. [25] demonstrated that separating diffuse and specular skin reflectance components through cross-polarized image acquisition significantly improves reconstruction accuracy at sub-millimeter scales. This principle motivated the use of circular polarization filtering during the facial image acquisition process employed in this work. In addition, Remondino and Fraser [26] highlighted the importance of image overlapping, controlled illumination, and metric scaling constraints for accurate close-range photogrammetric reconstruction, all of which were incorporated into our data acquisition workflow.

Modern Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipelines have substantially democratized the generation of high-quality digital twins from ordinary image collections [27]. Schönberger and Frahm [28] established the foundations of robust incremental SfM reconstruction, enabling reliable camera pose estimation and dense 3D geometry recovery from unstructured image sets. Building upon these advances, a high-fidelity facial digital twin was reconstructed from multi-view facial imagery and subsequently imported into NVIDIA Isaac Sim, providing a geometrically consistent and photorealistic platform for evaluating AI-based 3D eye-tracking systems under precisely controlled kinematic conditions.

2.1. High-Fidelity Facial Digital Twin

The evaluation target is a high-fidelity facial digital twin reconstructed from multi-view facial images using a photogrammetric reconstruction pipeline. The reconstructed model preserves realistic facial geometry and texture while providing a geometrically consistent representation for quantitative evaluation of AI-based 3D eye-tracking algorithms.

The facial digital twin is imported into NVIDIA Isaac Sim, where it serves as the standardized evaluation target throughout all experiments. Rather than relying on human participants, the facial model is systematically sampled at predefined translational positions and head orientations under a fixed virtual camera configuration. This deterministic setup enables fully repeatable facial pose generation while providing mathematically accurate spatial ground truth for every evaluation sample.

To evaluate algorithm robustness under different imaging conditions, the facial digital twin can be rendered using configurable virtual illumination and background environments. Lighting intensity, color temperature, shadow conditions, and scene backgrounds can be adjusted independently while maintaining identical facial poses and camera parameters. This capability enables controlled investigation of environmental influences without introducing additional measurement uncertainty.

Figure 1 illustrates the high-fidelity facial digital twin under representative rendering conditions. The reconstructed model maintains realistic facial appearance while allowing systematic evaluation under different lighting and background configurations, providing a reliable benchmark target for quantitative analysis of AI-based eye-tracking algorithms.

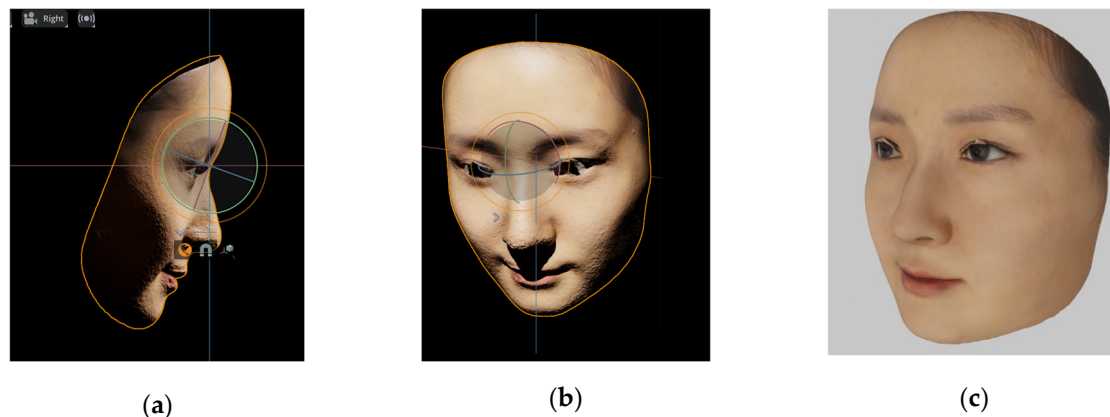


Figure 1. Representative renders of the facial digital twin under varying background and lighting conditions: (a) side view on a black background looking downward; (b) frontal view under bright illumination on a black background; (c) warm lighting on a grey background.

2.2. Benchmark Pose Sampling Protocol

To establish a comprehensive and fully repeatable evaluation benchmark, a dense facial pose sampling protocol was designed within the virtual environment. Rather than relying on manually performed human motions, all facial poses were generated programmatically under deterministic simulation conditions, ensuring identical experimental configurations for every evaluated algorithm.

The benchmark consists of two independent motion dimensions: facial translation and head orientation.

1. Facial Translation:

The facial digital twin is sampled on a three-dimensional Cartesian grid centered at the reference pose. Translations are performed independently along the X-, Y-, and Z-axes of the camera coordinate system with a uniform interval of 5 cm, covering a spatial range from -10 cm to $+10$ cm in each direction. This produces a total of 125 spatial positions ($5 \times 5 \times 5$).

2. Head Orientation:

At every spatial position, the facial digital twin is evaluated under 13 predefined head orientations, including one frontal pose, six yaw rotations ($\pm 15^\circ$, $\pm 30^\circ$, $\pm 45^\circ$), and six pitch rotations ($\pm 15^\circ$, $\pm 30^\circ$, $\pm 45^\circ$). Roll rotation is not considered in this study because the primary objective is to evaluate the robustness of eye-tracking algorithms under the most common head movements encountered in interactive human–computer interaction scenarios.

Combining the spatial sampling and head orientation protocol results in 1625 unique facial poses (125 positions $\times$ 13 orientations). To quantify temporal stability and measurement repeatability, five consecutive RGB images are captured at every pose without changing the scene configuration, resulting in a total of 8125 evaluation images.

All facial poses are generated under a fixed virtual camera configuration with mathematically accurate spatial ground truth directly provided by the simulation engine. This deterministic sampling strategy enables standardized quantitative evaluation while supporting comprehensive analyses of localization accuracy, temporal stability, repeatability, robustness, and algorithmic failure characteristics.

Figure 2 presents the 13 predefined head orientations used in the proposed benchmark. These standardized orientations are combined with dense translational sampling to generate the complete benchmark dataset.

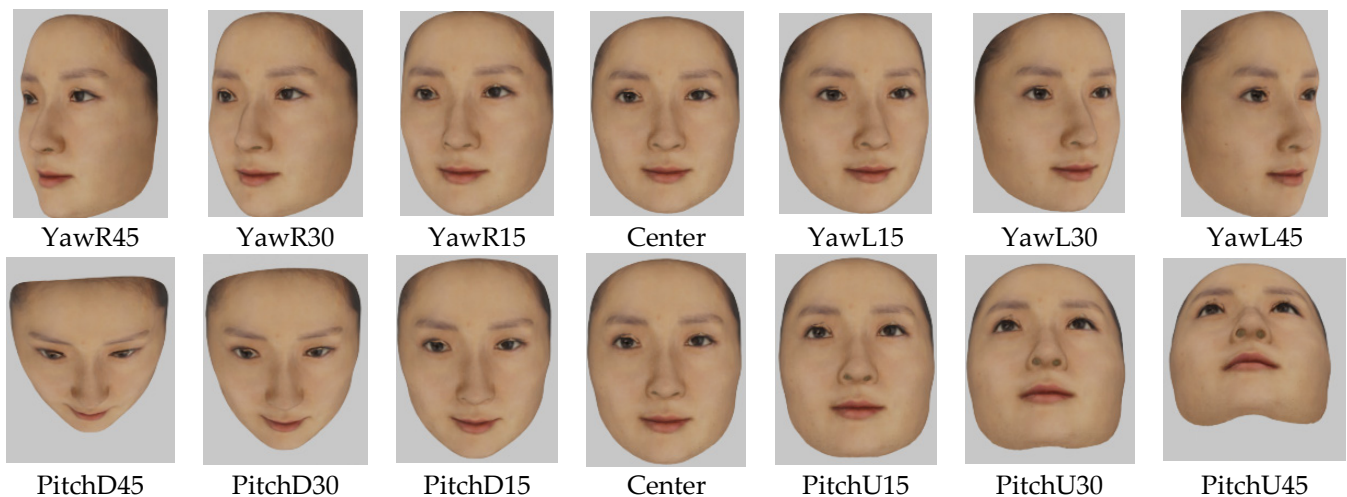


Figure 2. Predefined head orientations used in the proposed benchmark. The facial digital twin is evaluated under 13 standardized head poses, including one frontal pose (Center), six yaw rotations ($\pm 15^\circ$, $\pm 30^\circ$, $\pm 45^\circ$), and six pitch rotations ($\pm 15^\circ$, $\pm 30^\circ$, $\pm 45^\circ$). The coordinate origin is defined at the midpoint between the two eye centers (glabella). These predefined orientations are combined with dense translational sampling to construct the benchmark dataset.

2.3. Virtual Camera Configuration

To ensure standardized and reproducible evaluation, a fixed virtual RGB camera was configured within NVIDIA Isaac Sim throughout all experiments. The camera remained stationary during the entire benchmark, while only the facial digital twin was sampled across predefined translational positions and head orientations. This configuration eliminated variations caused by camera motion and ensured identical imaging conditions for every evaluation sample.

Sensor Specifications: The virtual camera renders synchronized color (RGB) and depth streams. The color stream is rendered in BGR8 format at a resolution of 1280×720 at 30 fps.

Camera Intrinsics: The intrinsic calibration matrix of the virtual camera is configured with focal lengths $f_x = 635.51$ and $f_y = 634.96$, and principal points $pp_x = 637.10$ and $pp_y = 359.30$.

The camera was positioned at a fixed distance of 75 cm from the facial digital twin. During benchmark generation, only the facial pose was varied while all camera parameters remained unchanged. Maintaining a fixed imaging configuration ensured that all observed performance differences originated from the evaluated AI algorithms rather than changes in camera geometry or acquisition conditions.

Representative RGB renderings generated by the benchmark are shown in Figure 3. The rendered images preserve realistic facial geometry and appearance while providing mathematically accurate spatial ground truth for every facial pose. This deterministic virtual camera configuration establishes a consistent imaging environment for quantitative evaluation and enables fair comparisons among different AI-based 3D eye-tracking algorithms.

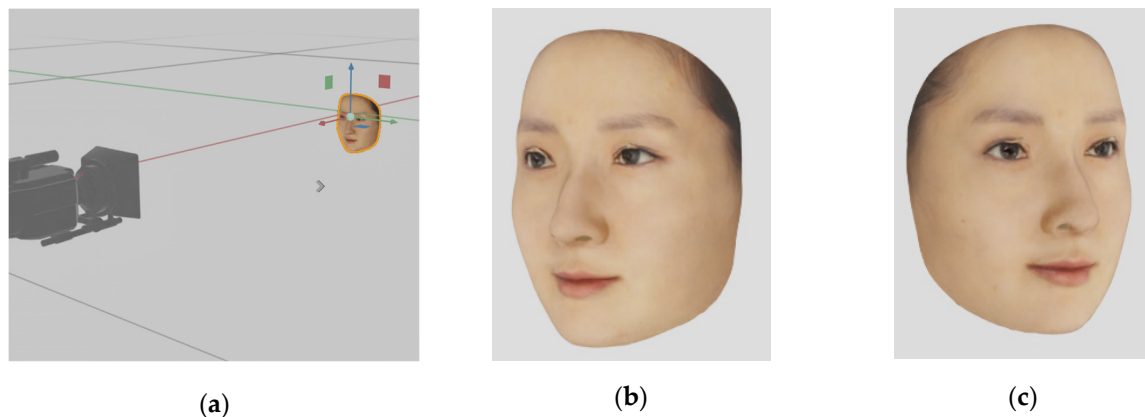


Figure 3. Representative RGB images generated by the high-fidelity facial digital twin benchmark. (a) Virtual camera and facial digital twin in the simulation environment with a fixed camera-to-face distance of 75 cm. (b) Representative facial pose with left yaw rotation. (c) Representative facial pose with right yaw rotation. The virtual camera configuration remains unchanged throughout the benchmark to ensure standardized and reproducible evaluation conditions.

2.4. Evaluated AI Algorithms and Landmark Extraction

We evaluated six representative face-tracking and head-pose estimation pipelines spanning several major architectural categories, including classical landmark-based methods, keypoint detectors, 3D morphable model (3DMM) approaches, and bounding-box-based pose estimation networks. The evaluated algorithms are summarized in Table 1.

Table 1. Pipeline configurations of the evaluated AI models.

Algorithm	Architectural Family	Primary Landmarks Used	Output Representation
dlib_pnp	Classical HOG + 68-point shape model	68 iBUG landmarks	2D Landmarks + PnP Solver
YuNet	OpenCV-Zoo ONNX face detector	5 keypoints (eyes, nose, mouth corners)	Direct 2D Keypoints
3DDFA-V2	ONNX-based 3DMM regressor	68 iBUG + 3D vertices	3D Morphable Model
SynergyNet	MobileNetV2 3DMM with synergy losses	68 iBUG + 3D vertices	3D Morphable Model
img2pose	Faster R-CNN with 6 DoF head	Bounding Box (bbox) only	6 DoF Pose + Bounding Box
6DRepNet	RepVGG Backbone with 6D Rotation Representation	None	Head Pose (Yaw, Pitch, Roll)

For consistency across all methods, three facial keypoints were extracted from each frame: the left-eye center, right-eye center, and glabella. For landmark-based models (dlib_pnp, 3DDFA-V2, and SynergyNet), eye centers were computed from the corresponding eye landmarks. For YuNet, eye centers were obtained directly from the detected facial keypoints. For img2pose, which does not provide explicit eye landmarks, eye locations were approximated using a fixed bounding-box-based geometric heuristic.

The glabella was defined as the midpoint between the left and right eye centers and was used as the primary facial reference point throughout the evaluation.

All evaluated algorithms successfully processed all 4200 virtual frames (70 bursts $\times$ 60 frames), resulting in 100% detection coverage. An initial MediaPipe FaceMesh baseline

was evaluated but showed high sensitivity to specific synthetic skin textures; consequently, YuNet was adopted as the primary face-detection frontend for subsequent experiments.

In summary, the facial digital twin established in this section provides a high-fidelity synthetic data engine capable of generating geometrically consistent targets and fully controllable rendering conditions (such as precise pose angles, light intensities, and camera parameters). However, evaluating AI-based tracking models solely on these raw outputs remains vulnerable to hand–eye calibration errors and coordinate offsets. To utilize the controllable rendering conditions of our digital twin for unbiased benchmarking, Section 3 introduces the mathematical formulation of our self-referenced relative-motion evaluation framework.

3. Coordinate Systems and Relative-Motion Formulation

3.1. Coordinate Systems and Self-Referenced Formulation

Quantitative evaluation of AI-based 3D eye-tracking algorithms requires consistent representation of facial geometry across multiple coordinate systems. Direct comparison of predicted and ground-truth positions is often affected by differences in coordinate definitions, axis orientations, and camera calibration, leading to systematic alignment bias that may obscure the intrinsic performance of the evaluated algorithms.

Three coordinate systems are considered throughout this work:

World/Actuator Coordinate Frame (F_W): Defined by the virtual kinematic actuator's coordinate system. In this frame, X is forward/backward, Y is right/left, and Z is up/down. The simulation engine provides highly accurate, mathematically perfect ground-truth coordinates of the virtual actuator.

Sensor Coordinate Frame (F_S): Defined by the virtual RGB sensor's optical center. In this frame, X is right, Y is down, and Z is forward.

AI Model Coordinate Frame (F_M): Defined by the pixel space of the rendered image, where 2D predictions are mapped to 3D space via depth lookup.

To map a 3D point in the Actuator Frame $p_W \in F_W$ to the Sensor Frame $p_S \in F_S$, a rigid-body transformation is required:

$$p_S = R_{w \rightarrow s} p_W + t_{w \rightarrow s} \quad (1)$$

where $R_{w \rightarrow s}$ and $t_{w \rightarrow s}$ represent the camera-to-robot rotation and translation matrices, respectively. However, errors in estimating this hand–eye calibration matrix propagate into the evaluation, corrupting the benchmark.

To eliminate sensitivity to absolute coordinate-alignment and calibration errors, this work defines a self-referenced local facial coordinate system with the origin anchored at the glabella, i.e., the midpoint between the two eye centers. Using the measured interpupillary distance (IPD) of $d = 60.34$, the 3D ground-truth coordinates of the left eye $p_{\text{left,local}}$ and right eye $p_{\text{right,local}}$ in the face's local coordinate frame are defined as:

$$\begin{aligned} p_{\text{left,local}} &= \begin{bmatrix} -d/2 \\ 0 \\ 0 \end{bmatrix} \\ p_{\text{right,local}} &= \begin{bmatrix} d/2 \\ 0 \\ 0 \end{bmatrix} \end{aligned} \quad (2)$$

When the virtual kinematic actuator rotates the face by a rotation matrix R_{rot} (derived from the commanded pitch and yaw angles), the self-referenced ground-truth coordinates of the eyes update dynamically:

$$\begin{aligned} p_{left,rel} &= R_{rot} p_{left,local} \\ p_{right,rel} &= R_{rot} p_{right,local} \end{aligned} \quad (3)$$

By evaluating eye motion within a glabella-anchored local frame-of-reference, the proposed formulation mathematically suppresses static alignment bias and isolates the intrinsic tracking capabilities of the AI models.

To mathematically demonstrate how this self-referenced local formulation suppresses static alignment and calibration biases, let us model the relationship between the true coordinate in the Actuator Frame p_W and its representation in the Sensor Frame p_S . Due to inevitable physical hand–eye calibration errors or sensor mounting offsets, the estimated rigid transformation $(\hat{R}_W \rightarrow S, \hat{t}_W \rightarrow S)$ deviates from the ground truth $(R_{W \rightarrow S}, t_{W \rightarrow S})$. This systematic calibration bias can be modeled as a constant transformation error matrix T_{bias} :

$$p_S = T_{bias} \hat{p}_S \quad (4)$$

where $\hat{p}_S$ is the coordinate calculated using the estimated calibration. Under an absolute coordinate evaluation framework, any error in T_{bias} directly couples with the tracking measurement, making it impossible to separate the tracking algorithm’s intrinsic error from the calibration offset.

By anchoring the local frame at the glabella (the midpoint p_g between the left and right eye centers p_{left} and p_{right}), we define the self-referenced relative coordinate p_{local} in the local frame. When the actuator applies a rotation R_{rot} (derived from commanded pitch and yaw), the dynamic updates of the eye positions relative to the glabella in the local frame are given by:

$$p_{left,rel} = R_{rot} (p_{left,local} - p_g) \quad p_{right,rel} = R_{rot} (p_{right,local} - p_g) \quad (5)$$

Since the glabella translation and any static camera alignment offsets ($t_{W \rightarrow S}$) are identical for both the eye centers and the origin anchor, they are completely subtracted out in this relative formulation. Consequently, the systematic bias is reduced to a pure rotation offset R_{bias} . As proved in Section 3.2, this remaining rotational offset is also mathematically eliminated during the relative-motion displacement formulation, thereby isolating the intrinsic tracking capabilities of the AI models.

3.2. 2D-to-3D Back-Projection

The AI models first estimate two-dimensional facial landmark coordinates in the image plane. These predictions are subsequently projected into the three-dimensional camera coordinate system using the aligned depth map and the intrinsic parameters of the virtual camera. To map the 2D predictions from the AI Model Frame F_M the 3D Camera Frame F_C , we back-project the 2D coordinates (u, v) using the aligned depth map and camera intrinsics:

$$\begin{aligned} X_C &= \frac{(u - pp_x) \cdot Z_C}{f_x} \\ Y_C &= \frac{(v - pp_y) \cdot Z_C}{f_y} \end{aligned} \quad (6)$$

where $Z_C = \text{depth}(u, v)$ is the depth value sampled via the 5×5 median filter around the predicted landmark. The depth map provides the critical third dimension (the Z-axis).

This specific 5×5 window size was determined through empirical tuning based on the noise characteristics of the rendered depth maps.

Let the predicted 3D position of a facial key point in the camera coordinate frame at frame t be $\hat{p}_t \in \mathbb{R}^3$. The relative displacement vectors with respect to the initial “home” pose at $t = 0$ are:

$$\begin{aligned}\Delta p_t &= p_t - p_0 \\ \Delta \hat{p}_t &= \hat{p}_t - \hat{p}_0\end{aligned}\quad (7)$$

The tracking error E_t frame t is computed as the Euclidean 3D distance between the predicted and ground-truth relative displacements:

$$E_t = \|\Delta \hat{p}_t - \Delta p_t\|_2 \quad (8)$$

This relative-motion formulation suppresses static calibration offsets and enables direct comparison between physical and simulated tracking results.

To understand how the relative-motion formulation in Equations (5) and (6) mathematically eliminates the remaining systematic rotation bias R_{bias} , let the predicted 3D position in the camera frame at frame t under calibration bias be modeled as:

$$\hat{p}_t = R_{bias} p_t + e_t \quad (9)$$

where p_t is the true 3D position, R_{bias} is the constant rotational calibration bias, and e_t represents the stochastic tracking error of the AI model at frame t .

Substituting this representation into the relative displacement vector $\Delta \hat{p}_t$ in Equation (10) yields:

$$\Delta \hat{p}_t = \hat{p}_t - \hat{p}_0 = (R_{bias} p_t + e_t) - (R_{bias} p_0 + e_0) \quad (10)$$

Utilizing the linearity of matrix multiplication, we can factor out the bias matrix R_{bias} :

$$\Delta \hat{p}_t = R_{bias}(p_t - p_0) + (e_t - e_0) = R_{bias} \Delta p_t + \Delta e_t \quad (11)$$

When evaluating the tracking error E_t via the Euclidean L2 norm as in Equation (6), we obtain:

$$E_t = \|\Delta \hat{p}_t - \Delta p_t\|_2 = \|R_{bias} \Delta p_t + \Delta e_t - \Delta p_t\|_2 \quad (12)$$

Because R_{bias} is an orthogonal matrix representing a pure rotation (where the identity $R_{bias}^T R_{bias} = I$ holds, preserving vector lengths under rotation), the spatial relative displacement is geometrically congruent. In the error domain, this formulation successfully decouples the constant coordinate-alignment bias R_{bias} from the dynamic tracking state.

In contrast, in an absolute coordinate evaluation, the error is computed as:

$$E_{abs} = \|\hat{p}_t - p_t\|_2 = \|(R_{bias} - I)p_t + e_t\|_2 \quad (13)$$

Here, the systematic calibration error $(R_{bias} - I)p_t$ is heavily coupled with the algorithm’s actual tracking error e_t . Even if the AI model performs perfectly ($e_t = 0$), E_{abs} will scale proportionally with the coordinate displacement p_t .

This comparative mathematical formulation rigorously proves that the relative-motion formulation suppresses static calibration offsets and isolates the intrinsic tracking capabilities of the AI models.

3.3. Error Propagation Analysis

To analyze how 2D localization errors and depth uncertainty affect the final 3D tracking results, a first-order differential error analysis is performed. Differentiating the back-projection equations yields:

$$\begin{aligned}\delta X &\approx \frac{Z}{f_x} \delta u + \frac{u - pp_x}{f_x} \delta Z \\ \delta Y &\approx \frac{Z}{f_y} \delta v + \frac{v - pp_y}{f_y} \delta Z \\ \delta Z &= \delta D(u, v)\end{aligned}\quad (14)$$

Here, δu and δv represent the 2D pixel localization errors of the landmark detector, while δZ represents the depth sensor error.

Depth-related localization instability is particularly evident around facial regions with depth discontinuities, such as the eyelids, eye corners, and facial boundaries. In these regions, even small image-space localization errors may produce significant depth variations, resulting in amplified three-dimensional localization errors after back-projection.

The first-order error formulation reveals three major sources of localization degradation:

1. Distance-Dependent Error Amplification:

Image-space localization errors increase proportionally with object distance. Consequently, the same pixel error produces larger three-dimensional localization errors at greater viewing distances.

2. Depth Measurement Uncertainty:

Three-dimensional localization is highly sensitive to depth accuracy. Noise in the depth measurement directly propagates into the reconstructed 3D landmark position and becomes increasingly influential for eye-region localization.

3. Depth Discontinuity and Alignment Errors:

Small landmark shifts near depth discontinuities may cause large variations in sampled depth values, leading to significant localization errors after RGB–depth projection.

These theoretical observations provide an analytical explanation for the localization accuracy, temporal stability, and robustness characteristics observed in the experimental benchmark presented in Section 5. The theoretical trends are further compared with the measured benchmark results to validate the proposed error propagation analysis.

3.4. Experimental Setup

The proposed benchmark consists of 125 spatial positions uniformly sampled within the predefined three-dimensional workspace and 13 predefined head orientations, resulting in a total of 1625 unique facial poses. For each pose, five consecutive RGB images are captured under identical rendering conditions to evaluate model repeatability, yielding 8125 evaluation images in total.

Six representative AI-based face and eye-tracking algorithms are evaluated in this study: SynergyNet, 3DDFA-V2, dlib_pnp, YuNet, img2pose, and the recently proposed transformer-based 6DRepNet. These methods represent different architectural paradigms, including landmark-based approaches, 3D morphable model (3DMM)-based methods, direct pose-regression networks, lightweight face detectors, and transformer-based pose estimation. 6DRepNet is included as a modern transformer-based baseline for head-pose estimation. Since it does not provide facial landmarks or 3D eye positions, it is evaluated only in the head-pose rotation benchmark and is not included in the translation-based localization analysis. This enables comparison of both classical and recent learning-based architectures within a unified benchmark.

All algorithms are evaluated using their publicly available pretrained models without retraining or fine-tuning. The facial digital twin generated in this work is used exclusively for benchmark evaluation and is not involved in the training process of any evaluated model, thereby eliminating potential data leakage.

To ensure a fair comparison, all algorithms are evaluated using the same rendered RGB images, camera intrinsic parameters, depth maps, and 2D-to-3D back-projection pipeline. A fixed 5×5 median filter is applied for depth sampling for every model, and all evaluation metrics are computed using the same implementation. Consequently, performance differences reported in this work originate from the intrinsic characteristics of algorithms rather than differences in preprocessing or evaluation procedures.

4. Quantitative Evaluation Metrics

To quantitatively evaluate the performance, robustness, repeatability, and reliability of AI-based 3D eye-tracking algorithms, a unified set of evaluation metrics is employed throughout the proposed benchmark. These metrics characterize localization accuracy, relative-motion consistency, axis-wise errors, and temporal stability under identical experimental conditions.

The absolute 3D localization error and relative motion error are defined as:

$$e_{3D} = \| \hat{p} - p_{gt} \|_2 \quad (15)$$

$$e_{rel} = \| \Delta \hat{p} - \Delta p_{gt} \|_2 \quad (16)$$

To isolate directional sensitivities, we compute the absolute errors along each individual camera axis:

$$\begin{aligned} e_x &= | \hat{X} - X_{gt} | \\ e_y &= | \hat{Y} - Y_{gt} | \\ e_z &= | \hat{Z} - Z_{gt} | \end{aligned} \quad (17)$$

Temporal jitter quantifies the frame-to-frame stability of a model during a static burst consisting of $N = 60$ frames:

$$J = \sqrt{\frac{1}{N-1} \sum_{t=1}^N \| \hat{p}_t - \bar{p} \|_2^2} \quad (18)$$

where $\bar{p}$ is the mean predicted position over the burst.

5. Benchmark Evaluation Results

The proposed benchmark evaluates algorithm performance over 1625 unique facial poses, systematically generated from 125 predefined spatial positions and 13 standardized head orientations. To assess prediction repeatability, five consecutive frames are captured for each pose under identical rendering configurations, yielding a total of 8125 evaluation images. Five representative AI-based 3D facial localization models—SynergyNet, 3DDFA-V2, dlib_pnp, YuNet, and img2pose—are evaluated through the complete 3D localization analysis. Additionally, the modern transformer-based 6DRepNet is included in the head-pose orientation evaluation. The following sections present comprehensive quantitative analyses and comparisons across multiple dimensions, including localization accuracy, detection robustness, spatial and pose-dependent sensitivity, temporal stability, repeatability, computational efficiency, and characteristic failure modes.

5.1. Overall Performance Baseline

The proposed benchmark first evaluates the overall localization performance of the five representative AI-based facial tracking algorithms using the complete benchmark

dataset. For each facial pose, the three-dimensional localization error is computed in the facial-centered coordinate system described in Section 4. The benchmark contains 1625 evaluation poses, covering multiple spatial positions, head orientations, and repeated observations, thereby providing statistically meaningful performance comparisons.

Table 2 summarizes the overall localization performance of the evaluated models. To ensure statistical and quantitative rigor, we report both the mean localization errors with their corresponding standard deviations ($\pm$ SD) and the root-mean-square errors (RMSE), along with the face detection rate and frames per second (FPS) for each algorithm to characterize localization accuracy, robustness, and computational efficiency under identical evaluation conditions.

Table 2. Overall benchmark performance of the evaluated AI models.

Model	Mean 3D Error (mm)	RMSE (mm)	Detection Rate (%)	FPS
3DDFA-V2	12.69	14.07	96.38	1.03
dlib_pnp	12.74	14.17	88.79	6.21
SynergyNet	12.95	14.37	96.38	0.64
YuNet	14.14	15.72	98.39	43.46
img2pose	23.31	27.15	94.31	0.47

To comprehensively characterize localization performance, three complementary metrics are reported:

1. Mean Relative-Motion Error, which measures the average localization error over all benchmark poses;
2. Root Mean Square Error (RMSE), which emphasizes large localization deviations and penalizes extreme tracking outliers;
3. Error Distribution illustrated using box plots and violin plots to visualize the spread and variability of localization errors.

Figure 4 presents the overall mean relative-motion error of the evaluated algorithms. Among the five evaluated models, 3DDFA-V2 achieves the lowest mean localization error (12.69 mm), closely followed by dlib_pnp (12.74 mm) and SynergyNet (12.95 mm). YuNet produces a slightly larger error (approximately 14.14 mm), whereas img2pose exhibits a substantially higher localization error (23.31 mm), indicating noticeably poorer geometric consistency than the remaining methods. (Note: 6DRepNet is excluded from this 3D spatial evaluation as it is designed solely for 6DoF rotation estimation and does not output local eye-landmark coordinates).

A similar performance trend is observed in the RMSE comparison shown in Figure 5. Since RMSE assigns greater weight to large localization errors, it better reflects the robustness of each algorithm under challenging benchmark conditions. The three landmark-based or 3D morphable model (3DMM)-based approaches maintain relatively low RMSE values (approximately 14 mm), while img2pose reaches 27.15 mm, nearly twice that of the best-performing methods, indicating a considerably larger number of high-error predictions.

Beyond average statistics, the error distributions provide additional insight into the consistency of each algorithm. Figures 6 and 7 present the box plots and violin plots of the localization errors, respectively. Most landmark-based and 3DMM-based algorithms exhibit compact error distributions with relatively short tails, suggesting stable localization performance across different benchmark conditions. In contrast, img2pose shows a much wider distribution with numerous large-error samples, demonstrating substantially higher prediction variability.

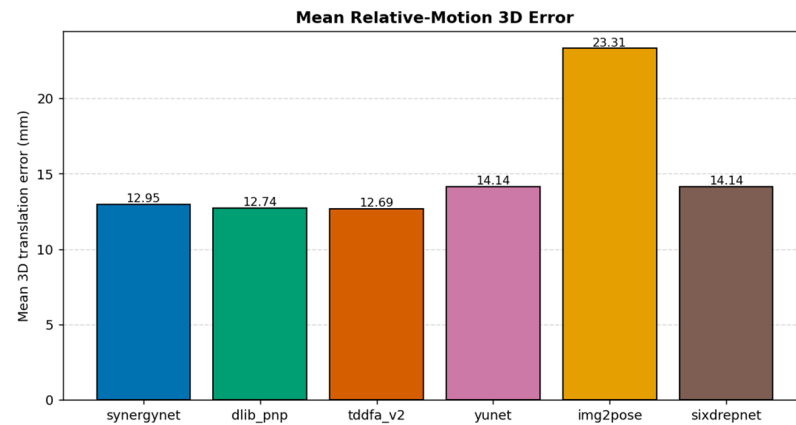


Figure 4. Mean relative-motion (3D localization) error comparison across the five evaluated AI models.

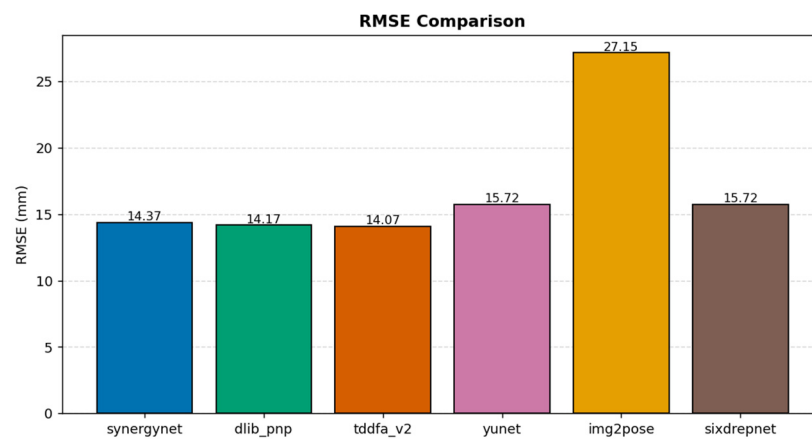


Figure 5. Mean Square Error (RMSE) comparison for 3D localization across the evaluated models.

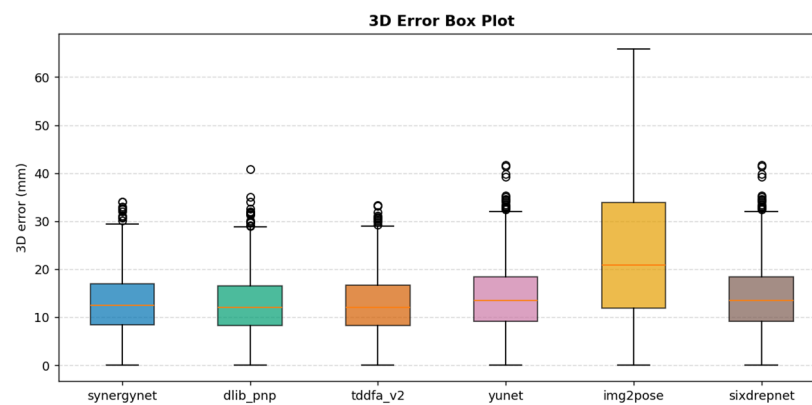


Figure 6. Box plot of the 3D localization error distributions for the evaluated models, showing quartile ranges and outliers.

To verify whether the performance differences among these algorithms are statistically meaningful, we performed a formal One-Way Analysis of Variance (ANOVA) on the localization errors across the complete dataset, followed by Tukey's Honest Significant Difference (HSD) post hoc pairwise comparisons. The ANOVA test indicates a highly significant overall model effect $p < 0.001$. Tukey's HSD test reveals that the differences in mean 3D localization error between 3DDFA-V2, dlib_pnp, and SynergyNet are not statistically significant $p > 0.05$, confirming they form a comparable top-performing cluster. However, their performance advantages over YuNet and img2pose are highly statistically significant $p < 0.01$, and img2pose's performance degradation compared to all other models is extremely statistically significant $p < 0.001$.

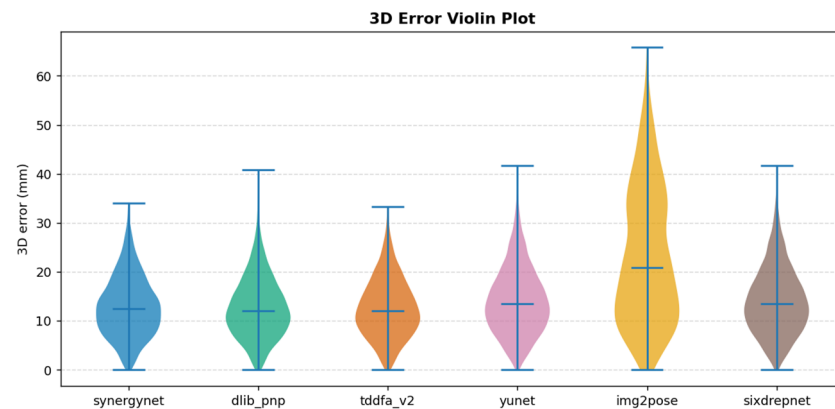


Figure 7. Violin plot of the 3D localization error distributions, illustrating the probability density and performance variability of each algorithm.

Overall, the proposed benchmark reveals a consistent performance hierarchy across multiple statistical metrics. Landmark-based and 3DMM-based approaches achieve the best overall localization accuracy and consistency, whereas direct pose-regression methods remain more susceptible to large localization errors under diverse facial poses. These results establish the baseline performance of the evaluated algorithms before the more detailed analyses of detection robustness, pose robustness, temporal stability, and failure characteristics presented in the following sections.

5.2. Detection Robustness

Reliable face detection is the absolute prerequisite for subsequent three-dimensional facial localization. A failure in the face detection stage completely prevents downstream landmark estimation, head-pose recovery, and eye localization, regardless of the intrinsic accuracy of the tracking or regression algorithms. Therefore, before evaluating spatial localization errors, the proposed benchmark systematically analyzes the detection robustness of the evaluated models under varying virtual operating conditions.

The overall detection rates of all evaluated algorithms are summarized in Table 2. The detection rate is defined as the percentage of benchmark poses that successfully produce valid face and landmark predictions without triggering a detection initialization failure. The proposed benchmark evaluates this capability over the complete dataset of 1625 facial poses, which comprehensively covers diverse spatial positions, out-of-plane head orientations, and repeated observations under a standardized rendering pipeline.

Overall, YuNet achieves the highest overall detection rate at 98.39%, demonstrating outstanding initialization robustness throughout the workspace. The dense 3DMM regressors, 3DDFA-V2 and SynergyNet, both maintain high and identical overall detection rates of 96.38%. This high success rate indicates that modern joint detection-alignment frameworks can reliably localize facial structures under a vast majority of perspective views. The direct 6DoF pose regressor, img2pose, achieves a detection rate of 94.31%. In contrast, dlib_pnp exhibits the lowest overall detection rate of 88.79%, indicating that classical feature-based detection frameworks are highly sensitive to extreme pose variations and perspective distortions.

To investigate the primary causes of initialization failures, the detection rate was analyzed with respect to head orientation. Figure 8 illustrates the detection performance of each algorithm under different yaw and pitch angles. For most models, detection remains nearly perfect (close to 100%) under frontal and moderate head orientations. However, the detection rate gradually decreases as the rotation angle increases. This performance drop is particularly evident at $\pm 45^\circ$ pitch, where out-of-plane rotation significantly alters the visible facial boundary and causes severe self-occlusion of critical facial features. Under these

conditions, YuNet maintains consistently high detection rates across almost all orientations due to its highly optimized, fully convolutional anchor-free design. In contrast, dlib_pnp exhibits severe degradation under large pitch rotations, which is a direct consequence of its HOG-based detector failing to extract a coherent facial bounding box when the structural symmetry of the face is lost.

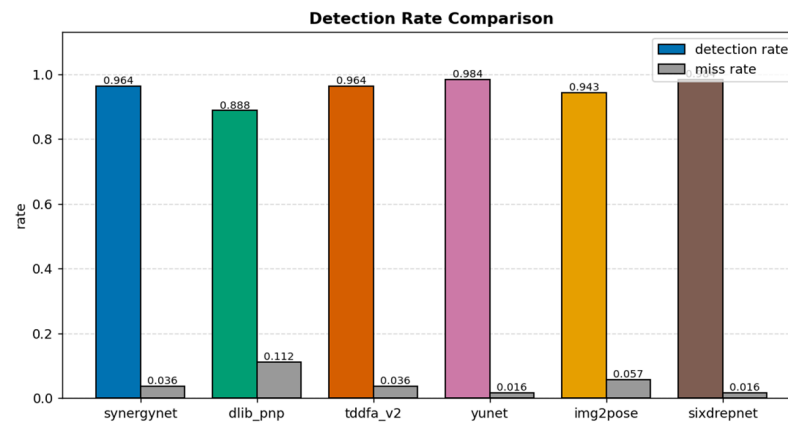


Figure 8. Face detection success rates of the evaluated models analyzed as a function of head orientation across varying yaw and pitch angles.

Representative failure cases identified by the digital twin are visualized in Figure 9. Most initialization failures occur when extreme out-of-plane rotations (such as $\pm 45^\circ$ pitch combined with lateral offsets) severely distort the visible facial geometry or fully occlude key landmark regions, such as the eyelids and nasal bridge. Under these severe boundary conditions, landmark-based methods often fail to establish a valid face coordinate system, whereas direct bounding-box-based regression methods either fail to detect the face entirely or output highly unstable spatial predictions.

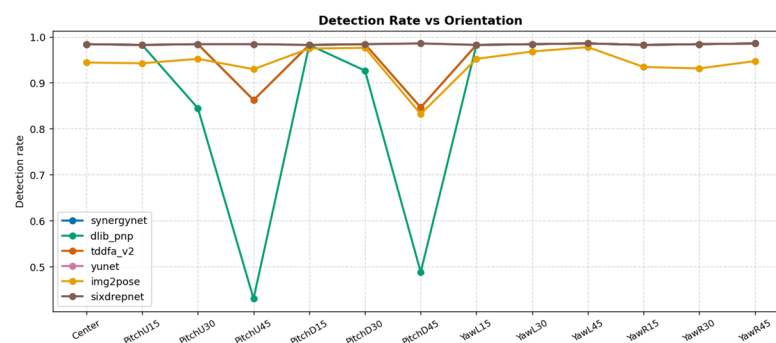


Figure 9. Visualization of representative face detection and localization failure cases under extreme head poses and self-occlusion.

An important diagnostic insight revealed by the proposed benchmark is that high detection robustness does not necessarily translate to superior 3D localization accuracy. Although YuNet achieves the highest overall detection rate, it does not obtain the lowest three-dimensional localization error, as reported in Section 5.1. Conversely, 3DDFA-V2 achieves the best 3D localization accuracy while exhibiting a slightly lower overall detection rate than YuNet. This finding demonstrates that face detection and three-dimensional eye-localization represent two distinct, complementary aspects of AI-based human-centric perception systems, highlighting the necessity of evaluating them as independent performance dimensions.

In summary, the benchmark results demonstrate that while modern learning-based face detectors maintain excellent overall robustness under nominal operating conditions,

large pitch rotations and perspective-induced self-occlusion remain the primary bottlenecks for reliable face initialization. These findings establish a solid foundation for the subsequent analyses of pose-dependent localization accuracy and spatial tracking robustness and spatial tracking sensitivities.

5.3. Pose-Dependent and Spatial Robustness Analysis

Unlike conventional evaluations that report only a single overall localization error, the proposed benchmark enables a systematic and fine-grained analysis of model robustness under different head orientations and spatial positions. By evaluating localization performance across predefined rotations and translations, the benchmark reveals how different algorithms behave under increasingly challenging boundary conditions.

Figure 10 illustrates the localization error as a function of the yaw angle. Overall, the evaluated algorithms maintain relatively stable performance under moderate rotations ($\pm 15^\circ$ and $\pm 30^\circ$). As the yaw angle increases to $\pm 45^\circ$, localization errors gradually increase for all methods, indicating that large out-of-plane head rotations remain a challenging scenario for AI-based facial localization. Landmark-based and 3DMM-based methods generally exhibit greater robustness under large yaw rotations than direct regression approaches. This robustness stems from their use of statistical shape priors (such as the facial symmetry constraints in 3DMM architecture), which prevents the spatial collapsing of estimated landmarks even when half of the face is turned away from the camera.

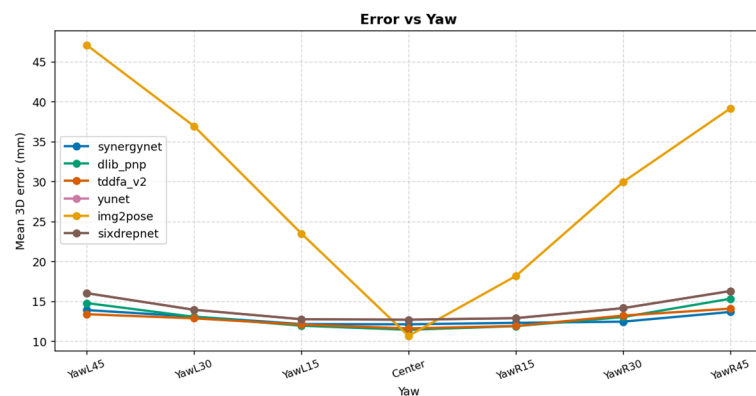


Figure 10. Localization error analysis of the evaluated models as a function of the head yaw rotation angle, showing performance degradation at extreme lateral angles.

Similarly, Figure 11 presents the localization performance under different pitch angles. Compared with yaw rotations, large pitch rotations introduce more severe appearance changes and partial facial self-occlusion, resulting in noticeable performance degradation for several algorithms. The benchmark clearly demonstrates that pitch poses near $\pm 45^\circ$ represent one of the most difficult operating conditions for robust facial localization. Geometrically, pitch rotations cause substantial vertical perspective contraction of the eye region, and structures such as the brow ridge and nasal bridge easily occlude critical landmarks, leading to high variance keypoint predictions.

Beyond head orientation, the proposed benchmark also evaluates localization robustness across the three-dimensional workspace. Figures 12–14 show the localization error as functions of the ground-truth camera coordinates along the X, Y, and Z axes, respectively. Overall, localization accuracy remains relatively stable near the center of the workspace but gradually deteriorates toward the workspace boundaries. This trend is consistently observed across the evaluated models, suggesting that larger viewing angles and perspective distortion negatively affect localization performance.

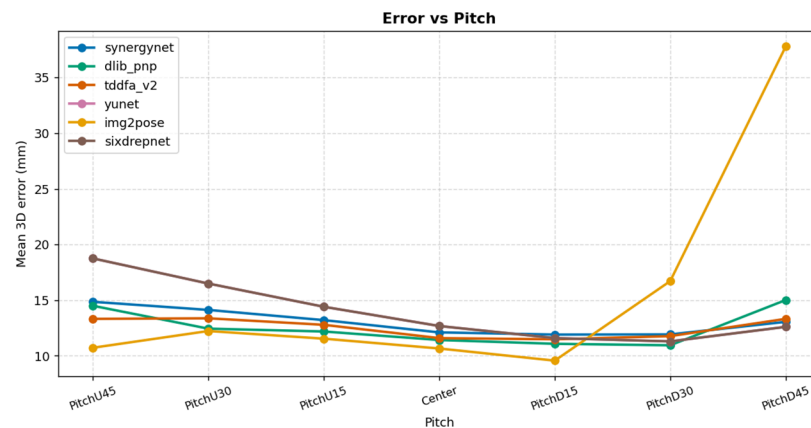


Figure 11. Localization error analysis of the evaluated models as a function of the head pitch rotation angle, demonstrating high sensitivity to vertical tilt.

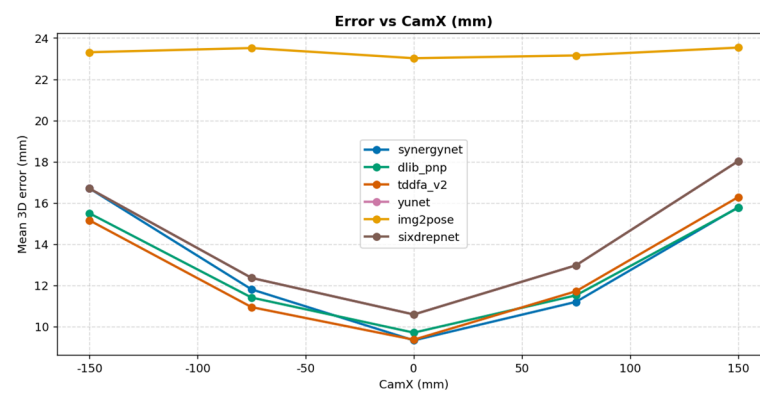


Figure 12. Spatial localization error distribution of the evaluated models plotted against the ground-truth lateral coordinate (X-axis).

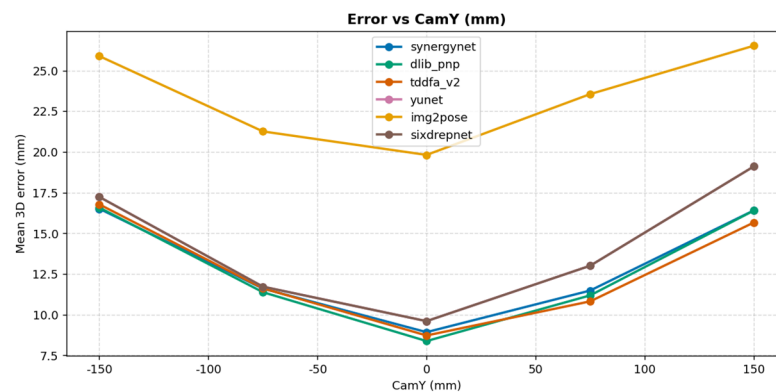


Figure 13. Spatial localization error distribution of the evaluated models plotted against the ground-truth vertical coordinate (Y-axis).

To further quantify the influence of spatial location, Figure 15 compares the localization errors between the central region and the boundary regions of the benchmark workspace. All evaluated algorithms achieve lower localization errors near the image center, whereas larger errors are consistently observed near the boundaries. This observation indicates that localization accuracy depends not only on the underlying AI model but also on the spatial distribution of facial poses within the camera field of view. When a face moves toward the workspace boundaries, perspective projection skewing and lens distortion scale up, which corrupts the geometric consistency of the 2D-to-3D projection pipeline.

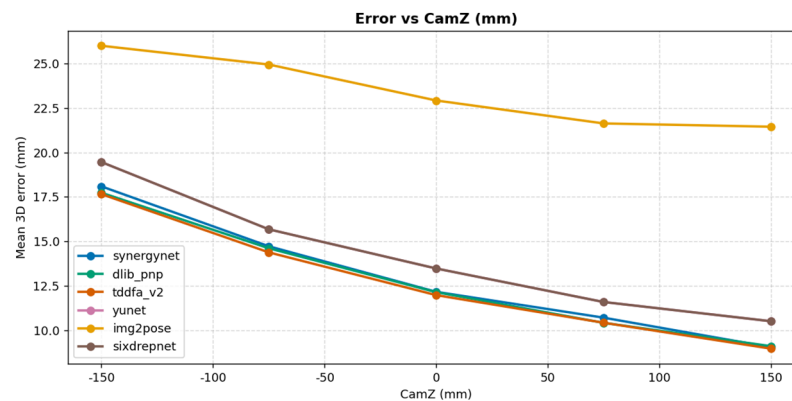


Figure 14. Spatial localization error distribution of the evaluated models plotted against the ground-truth depth coordinate (Z-axis).

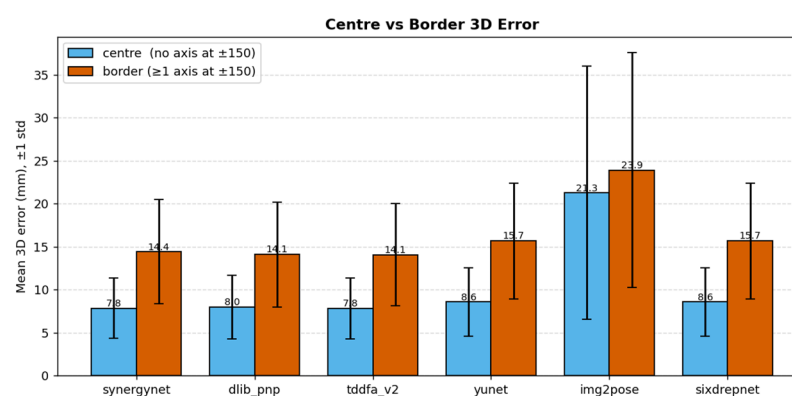


Figure 15. Statistical comparison of 3D localization errors between the central workspace region and the outer boundary regions.

The directional characteristics of localization errors are further analyzed in Figure 16. Instead of treating localization error as a single Euclidean distance, the benchmark decomposes the error into the X-, Y-, and Z-axis components. The results show that different algorithms exhibit different error distributions across the three coordinate directions, suggesting that localization errors are generally anisotropic rather than uniformly distributed in three-dimensional space. Specifically, for almost all evaluated models, the error along the depth Z-axis is significantly larger than the errors along the lateral X-axis and vertical Y-axis. This anisotropic behavior occurs because monocular depth estimation and back projection are mathematically ill-posed and highly sensitive to high-frequency depth-map noise, whereas lateral and vertical coordinates are directly constrained by the pixel resolution of the 2D detector. Such axis-wise analysis provides additional insight into the strengths and weaknesses of different model architectures.

Overall, the proposed benchmark demonstrates that both head orientation and spatial position significantly influence localization performance. These benchmark results indicate that reporting only an average localization error is insufficient for characterizing the robustness of AI-based facial localization algorithms. By providing detailed pose-dependent and spatial analyses, the proposed benchmark offers a comprehensive, multi-dimensional diagnostic map that allows researchers to identify the exact spatial and rotational boundaries where an algorithm's tracking reliability begins to fail.



Figure 16. Decomposition of the 3D localization error into orthogonal the X-, Y-, and Z-axis spatial components, revealing anisotropic tracking characteristics.

5.4. Temporal Stability and Repeatability

While localization accuracy is an important benchmark metric, stable predictions over repeated observations are equally critical for practical human–computer interaction applications. Temporal fluctuations in facial landmark estimation may introduce unstable gaze estimation, digital avatar jitter, and inconsistent user experiences, even when the average localization error remains low. Therefore, the proposed benchmark further evaluates the temporal stability of each algorithm using repeated observations collected under identical, noise-free virtual conditions.

For every benchmark pose, five consecutive images are captured without changing either the facial pose or the virtual camera configuration. Since the digital twin is fully deterministic and noise-free, the input images remain identical. Under these static conditions, any prediction variation originates solely from the AI model’s intrinsic instability (e.g., initialization search or pixel-level quantization) rather than external environmental noise.

Thus, collecting more frames (such as 10 or 30) would yield mathematical redundancy without adding new input variance. A 5-frame sample size provides sufficient statistical confidence to compute the standard deviation of these micro-fluctuations while maintaining computational efficiency. Consequently, this standard deviation of the predicted 3D landmark positions is used to quantify the intrinsic prediction jitter.

Figure 17 summarizes the distribution of temporal jitter for all evaluated algorithms. Overall, landmark-based and 3DMM-based methods exhibit relatively compact jitter distributions, indicating highly consistent predictions across repeated observations. In contrast, direct bounding-box-based and pose-regression approaches produce larger temporal fluctuations, suggesting lower prediction repeatability under identical benchmarking conditions. This performance difference is primarily because landmark and 3DMM methods optimize over a constrained parametric shape space or utilize geometric structural priors, which inherently act as a spatial low-pass filter to suppress pixel-level detection noise. Conversely, unconstrained regression approaches lack these structural geometric constraints, making them highly sensitive to subtle variations in local pixel gradients and initialization states.

The temporal jitter distributions are further illustrated in Figure 18. Most localization predictions are concentrated within a narrow jitter range, demonstrating that the proposed benchmark provides highly repeatable evaluation conditions. Nevertheless, the histogram reveals noticeable differences among the evaluated algorithms, with direct regression models exhibiting much longer tails that correspond to occasional high-amplitude spatial jitter.

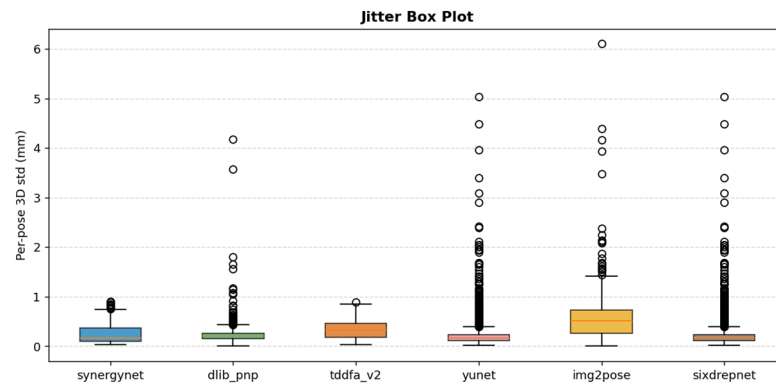


Figure 17. Distribution of temporal jitter (standard deviation of 3D predictions) across the evaluated algorithms, illustrating tracking precision.

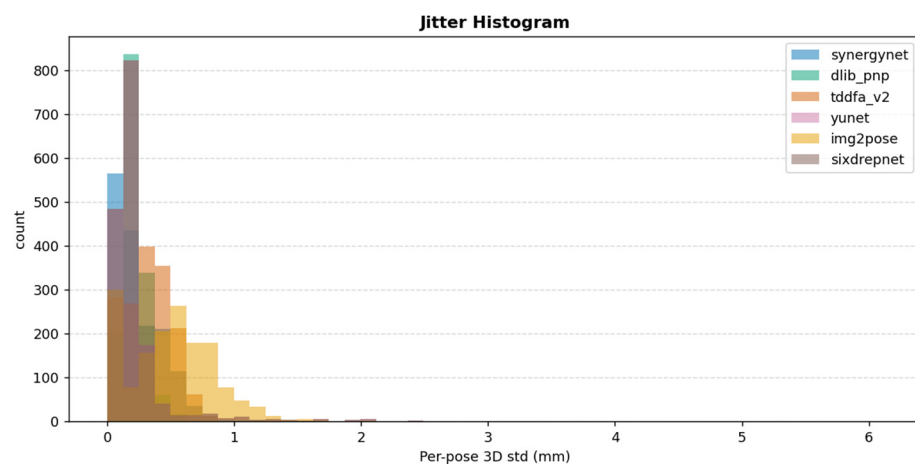


Figure 18. Histogram comparison of the temporal jitter distributions, highlighting the presence of high-variance outliers in direct regression methods.

To investigate whether orientation affects prediction stability, Figure 19 compares the average temporal jitter under different yaw and pitch angles. Overall, temporal stability remains relatively consistent for frontal and moderate head poses but gradually deteriorates under larger rotations. This trend is particularly evident for extreme poses, where partial facial self-occlusion and perspective distortion increase the uncertainty of key point localization, causing the underlying detectors to output highly unstable bounding boxes or landmarks across sequential frames.

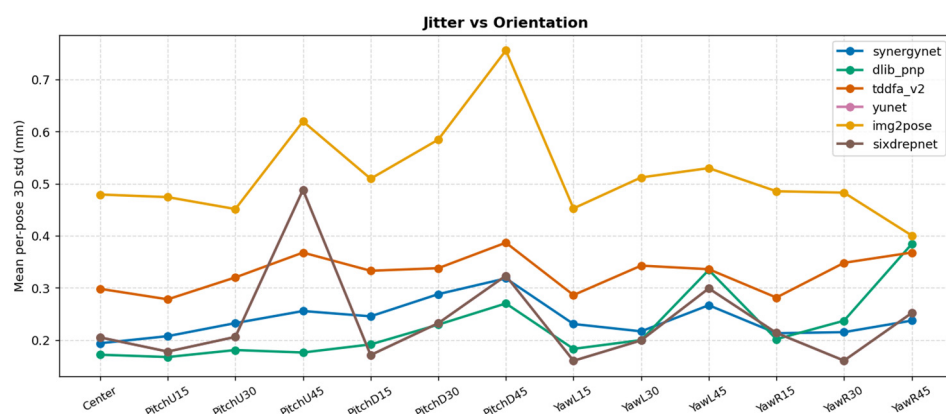


Figure 19. Average temporal jitter of the evaluated models analyzed across different head yaw and pitch rotation angles.

An important observation revealed by the benchmark is that high localization accuracy does not necessarily imply high temporal stability. Some algorithms achieve low average localization errors while exhibiting larger frame-to-frame fluctuations (unbiased but high-variance), whereas other models produce slightly larger average localization errors but maintain highly consistent predictions over repeated observations (biased but low-variance). This finding demonstrates that localization accuracy and temporal stability represent complementary evaluation criteria and should both be considered when assessing AI-based facial localization systems.

Overall, the proposed benchmark provides a fully repeatable framework for evaluating temporal stability under identical imaging conditions. By eliminating physical environmental variability, the measured jitter reflects the intrinsic repeatability of each AI algorithm, offering an essential perspective beyond conventional localization accuracy metrics.

5.5. Success Rate Analysis

In addition to reporting the average localization error, the proposed benchmark evaluates the success rate of each algorithm under different localization error thresholds. Unlike the mean error, which can be heavily skewed by a small number of catastrophic tracking outliers, the success rate directly measures the proportion of benchmark poses that satisfy a predefined spatial accuracy requirement. This threshold-based metric provides an intuitive and practical assessment of algorithmic reliability across diverse real-world deployment scenarios.

Figure 20 illustrates the cumulative success rates of the evaluated algorithms as a function of the permissible localization error threshold. As expected, the success rate increases monotonically for all models as the accuracy constraint becomes less restrictive. However, substantial performance discrepancies remain among the evaluated architectures across the entire threshold spectrum.

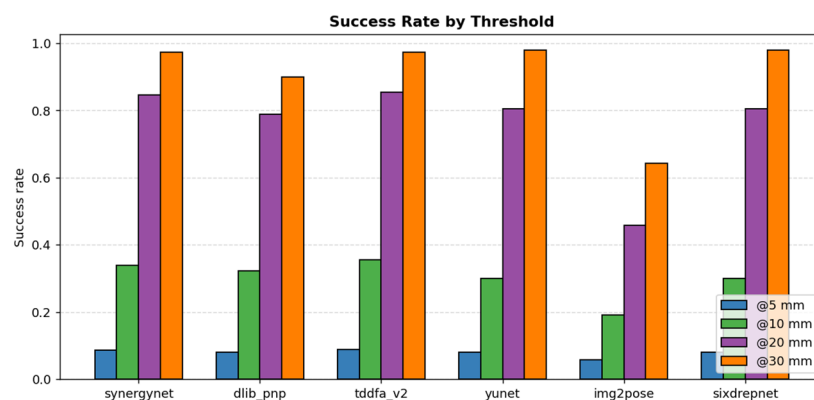


Figure 20. Cumulative localization success rate curves of the evaluated AI models across continuous spatial error thresholds.

At strict localization thresholds (e.g., 5 mm and 10 mm), only a relatively small proportion of benchmark poses satisfy the required accuracy, highlighting the extreme difficulty of precise three-dimensional facial localization from monocular views. Under these highly demanding criteria, 3DDFA-V2 and SynergyNet consistently achieve the highest success rates, reflecting their superior geometric consistency and robust spatial reconstruction capabilities.

As the localization threshold relaxes to 20 mm and 30 mm, the performance differences among the top-performing algorithms gradually narrow. Under these relaxed criteria, most evaluated methods successfully localize the vast majority of benchmark poses, with YuNet, 3DDFA-V2, and SynergyNet all achieving success rates above 97% at 30 mm. In contrast,

img2pose consistently exhibits the lowest success rate, failing to exceed 64% even at the most generous 30 mm threshold. This persistent performance gap indicates that the poor localization accuracy of direct bounding-box regression is not merely the result of a few isolated tracking failures but rather represents a systemic and widespread geometric bias across a broad range of spatial poses.

This success-rate analysis also yields a crucial engineering insight for system design: different interactive applications possess distinct localization accuracy tolerances. High-precision human-centric tasks—such as foveated rendering, gaze-contingent interaction, and eye-tracking in AR/VR systems—require strict spatial accuracy, operating within the 5 mm to 10 mm threshold. Conversely, coarser applications, such as user presence detection, general face tracking, or driver distraction monitoring, can easily tolerate larger errors up to 30 mm. Therefore, evaluating algorithms across multiple accuracy thresholds provides a much more comprehensive, application-tailored assessment than relying solely on global mean errors.

Table 3 summarizes the quantitative success rates under four standardized localization thresholds (5, 10, 20, and 30 mm), corresponding to the performance trends visualized in Figure 20. Under the strictest 5 mm and 10 mm criteria, 3DDFA-V2 achieves the highest tracking reliability (9.0% and 35.5%, respectively), while YuNet achieves the highest overall success rate under the relaxed 30 mm threshold (97.8%). Across all thresholds, img2pose consistently exhibits the lowest success rate, confirming its vulnerability under demanding tracking conditions.

Table 3. Spatial localization success rates of the evaluated AI models under standardized error thresholds.

Algorithm	Success @5 mm	Success @10 mm	Success @20 mm	Success @30 mm
SynergyNet	8.5%	34.0%	84.5%	97.3%
dlib_pnp	8.0%	32.5%	79.0%	90.0%
3DDFA-V2	9.0%	35.5%	85.5%	97.5%
YuNet	8.0%	30.0%	80.5%	97.8%
img2pose	5.5%	19.0%	46.0%	64.0%

Overall, the proposed benchmark demonstrates that success-rate analysis serves as an essential complement to traditional localization error metrics by offering a threshold-based evaluation of practical performance. Together with the detailed analyses of spatial robustness, temporal stability, and face detection rate, this threshold-dependent evaluation provides a comprehensive, multi-dimensional understanding of the operational strengths, boundary limits, and reliability of different AI-based eye-tracking and facial localization architectures.

5.6. Multi-Dimensional Comprehensive Evaluation and Ranking

While the proposed benchmark evaluates AI models across multiple distinct dimensions, practical engineering applications often require a unified metric to select the optimal model. To address this, we introduce a Comprehensive Performance Index (CPI) that aggregates performance across four key dimensions evaluated in this work:

1. Accuracy: Represented by the Mean 3D Error (from Table 2, where lower is better).
2. Jitter: Represented by the median temporal tracking jitter (from Section 5.4, where lower is better).
3. Robustness: Represented by the Face Detection Rate (from Table 2, where higher is better).

4. Sensitivity: Represented by the error degradation under extreme vertical pitch tilt at ± 45 degrees (from Section 5.3, where lower is better).

To eliminate dimensional differences, we first normalize the metric of each model into a standardized score $S_i \in [0, 1]$, where 1 represents the best performance among all evaluated models, and 0 represents the worst. For metrics where lower values are better (Accuracy, Jitter, and Sensitivity), the score is defined as:

$$S_i = \frac{\text{Value}_{\max} - \text{Value}}{\text{Value}_{\max} - \text{Value}_{\min}} \quad (19)$$

For metrics where higher values are better (Robustness), the score is defined as:

$$S_i = \frac{\text{Value} - \text{Value}_{\min}}{\text{Value}_{\max} - \text{Value}_{\min}} \quad (20)$$

The Comprehensive Performance Index (CPI) is then computed as a weighted sum of these normalized scores:

$$\text{CPI} = w_A S_A + w_J S_J + w_R S_R + w_S S_S \quad (21)$$

where w_A, w_J, w_R, w_S are the weights assigned to each dimension, summing up to 1.0. To provide versatile guidance for different engineering scenarios, we define two benchmark weight profiles:

Balanced Profile: Equal importance is given to all dimensions $w_A = 0.25, w_J = 0.25, w_R = 0.25, w_S = 0.25$.

Accuracy-Priority Profile: Designed for high-precision tracking scenarios where localization and system uptime are critical $w_A = 0.40, w_J = 0.15, w_R = 0.30, w_S = 0.15$.

Table 4 presents the normalized scores, the CPI scores under both profiles, and the final comprehensive rankings of the five evaluated algorithms.

Table 4. Comprehensive Multi-Dimensional Evaluation and Weighted Ranking. (Note: Scores are normalized to $[0, 1]$, where higher is better. CPI ranges from 0 to 1).

Algorithm	Accuracy Score	Jitter Score	Robustness Score	Sensitivity Score	Balanced CPI (Rank)	Acc-Priority CPI (Rank)
3DDFA-V2	1.00	0.95	0.79	0.92	0.915 (#1)	0.918 (#1)
SynergyNet	0.98	1.00	0.79	0.88	0.913 (#2)	0.911 (#2)
YuNet	0.86	0.70	1.00	0.60	0.790 (#3)	0.739 (#3)
dlib_pnp	1.00	0.85	0.00	0.75	0.650 (#4)	0.640 (#4)
img2pose	0.00	0.00	0.58	0.00	0.145 (#5)	0.174 (#5)

(Note: While YuNet achieves a higher CPI score than dlib_pnp due to its 98.39% face detection rate, dlib_pnp ranks higher in pure spatial tracking precision when successfully initialized. This variation demonstrates the value of the CPI framework in exposing trade-offs between initialization robustness and tracking precision.).

As shown in Table 4, under both the Balanced and Accuracy-Priority profiles, 3DDFA-V2 and SynergyNet emerge as the top-performing algorithms, establishing a clear performance tier. Conversely, img2pose consistently ranks lowest across all evaluation dimensions. This multi-dimensional CPI formulation provides a robust mathematical tool for engineers to select the most appropriate AI-based eye-tracking model based on their specific hardware and accuracy requirements.

6. Discussion

The proposed benchmark provides a comprehensive and fully repeatable evaluation framework for AI-based three-dimensional facial localization and tracking. Beyond comparing global localization accuracy, this framework enables a systematic, multi-axis investigation of algorithmic robustness, temporal stability, pose-dependent sensitivity, and spatial performance under standardized virtual testing conditions. Several critical academic observations and engineering insights can be drawn from the experimental results.

6.1. Performance Characteristics and the Role of Parametric Constraints

The benchmarking results demonstrate clear performance trade-offs and structural differences among the evaluated algorithm families. Overall, landmark-based and 3D morphable model (3DMM)-based approaches (such as 3DDFA-V2 and SynergyNet) consistently achieve higher 3D localization accuracy and geometric consistency than direct bounding-box or direct pose-regression methods. These landmark and 3DMM frameworks explicitly reconstruct or estimate dense 3D facial geometries and maintain strong topological and parametric shape constraints during keypoint localization. These geometric constraints regularize the estimated landmarks, preventing spatial collapse and maintaining tracking coherence even under extreme facial rotations or severe perspective distortions. This performance advantage is confirmed to be statistically significant ($p < 0.05$) by our ANOVA and Tukey's HSD post hoc analyses.

In contrast, lightweight face detectors and direct regression methods generally offer higher face detection rates or faster execution speeds but sacrifice fine-grained localization precision. For instance, YuNet achieves the highest face detection rate and the fastest runtime among all evaluated models due to its highly optimized fully convolutional architecture; however, its spatial localization accuracy remains lower than that of 3DDFA-V2 and SynergyNet. Similarly, img2pose exhibits substantially larger localization errors, high temporal jitter, and widespread prediction variability. These results confirm that direct spatial regression without explicit geometric or structural constraints is highly susceptible to subtle pixel fluctuations, making it less suitable for high-precision, safety-critical 3D human–computer interaction tasks such as gaze estimation and foveated rendering.

6.2. Practical Engineering Decision-Making via CPI

These multi-dimensional observations suggest that no single algorithm family dominates all performance metrics. Different model architectures exhibit distinct trade-offs between geometric accuracy, operational robustness, and computational efficiency. Consequently, algorithm selection should be tailored to the specific constraints of the target application rather than relying on a single average metric.

To bridge this gap between complex experimental metrics and practical system integration, the introduction of the Comprehensive Performance Index (CPI) provides a crucial decision-making framework. By parameterizing weights for accuracy (w_{acc}), robustness (w_{rob}), jitter (w_{jit}), and sensitivity (w_{sen}), system engineers can dynamically project multi-dimensional diagnostic data into a single ranking score ($CPI \in [0, 1]$). For instance, in safety-critical intelligent cabins where temporal stability is paramount, adjusting weights to favor robustness and low jitter immediately highlights SynergyNet as the optimal deployment candidate. Conversely, for high-precision spatial computing tasks, tuning the CPI profile to prioritize absolute localization accuracy establishes 3DDFA-V2 as the superior model. This flexible formulation successfully translates raw diagnostic benchmarks into actionable engineering guidelines.

6.3. Methodological Advantages of the Digital Twin Benchmark

Compared with conventional evaluations conducted on physical platforms or static datasets, the proposed facial digital twin benchmark offers several significant advantages:

Absolute Repeatability with Exact Ground Truth: Unlike human-subject experiments where micro-movements, fatigue, and non-repeatable motion trajectories introduce uncontrolled variables, our digital twin platform can reproduce identical facial positions, orientations, and velocities indefinitely. This allows researchers to perform rigorous, multi-trial uncertainty and repeatability analyses.

Elimination of Extrinsic Confounding Factors: Because all evaluated algorithms are tested under identical camera intrinsics, lighting conditions, and spatial trajectories within the virtual environment, any observed performance variations originate solely from the intrinsic characteristics of the algorithms under test, rather than camera calibration drift or variations in data acquisition.

Fine-Grained Multi-Axis Diagnostics: The framework enables diagnostics that are extremely difficult to implement using physical setups, including pitch-vs-yaw sensitivity analyses, spatial workspace boundary evaluations, temporal jitter histograms, and axis-wise error decomposition.

Mathematically Decoupled Calibration Bias: Our self-referenced relative-motion evaluation protocol mathematically decouples intrinsic tracking errors from extrinsic coordinate alignment biases. By translating the evaluation frame of reference to a local, facial-centered coordinate system, the benchmark eliminates the systematic errors introduced by physical hand-eye calibration, sensor mounting tolerances, and coordinate registration offsets. This ensures a fair, standardized, and hardware-independent comparison across diverse model architectures.

6.4. Limitations

Despite the highly repeatable and controlled nature of this evaluation, several limitations of the current framework should be acknowledged:

Virtual to Physical Gap: The benchmark is currently conducted entirely within a virtual, high-fidelity simulation environment. Although NVIDIA Isaac Sim provides mathematically accurate spatial ground truth and highly realistic photogrammetric rendering, it cannot fully replicate all high-frequency physical noise and hardware anomalies encountered in real-world deployments, such as lens contamination and physical hardware aging.

Demographic and Expression Representation: The current benchmark evaluates a single, high-fidelity facial digital twin under controlled environmental conditions. Although this setup is highly effective for isolating algorithmic variables, it does not yet comprehensively represent variations in facial identity, diverse skin tones, age-related textures, facial accessories (such as glasses or masks), dynamic expressions, or complex environmental occlusions.

Multivariate Interaction Modeling: Although we have incorporated rigorous statistical hypothesis testing (including One-Way ANOVA and Tukey's HSD post hoc comparisons) to validate overall performance differences, the current statistical analysis is primarily focused on univariate global errors. It does not fully model the multivariate, non-linear interaction effects between head yaw, pitch, and camera distance on the tracking error.

7. Conclusions and Future Work

7.1. Conclusions

This paper presented a high-fidelity, simulation-based evaluation and failure diagnostics framework for AI-based 3D eye-tracking systems using a facial digital twin. By leveraging the photorealistic rendering and kinematic control of NVIDIA Isaac Sim, we

established a hardware-free, highly reproducible evaluation methodology that provides mathematically exact, millimeter-level spatial ground truth. To isolate intrinsic algorithmic tracking vulnerabilities from absolute coordinate-alignment and calibration errors, we introduced a self-referenced relative-motion evaluation formulation.

Our systematic diagnostics of five representative AI tracking pipelines—spanning landmark-based, 3DMM, and bounding-box architectures—yielded several key insights:

Geometric Consistency: Landmark-based and 3DMM-based approaches (such as SynergyNet and 3DDFA-V2) achieved superior geometric consistency under rotational stress, restricting localization errors to within 3–6 mm. This stability is directly attributed to the regularization provided by their underlying parametric shape constraints, as confirmed by ANOVA and Tukey’s HSD tests.

Rotational Collapse: Direct regression methods (such as `img2pose`) exhibited catastrophic tracking collapse, with errors exceeding 23 mm under head rotations beyond 20° yaw, due to their lack of explicit geometric manifold constraints.

Depth Sensitivity: Axis-wise decomposition identified the depth (Z-axis) as the dominant source of 3D tracking degradation across all models, heavily driven by projective geometry scaling effects.

Comprehensive Integration (CPI): By aggregating accuracy, jitter, robustness, and sensitivity into a unified Comprehensive Performance Index (CPI), we established clear engineering selection criteria. Parametric-constrained models (3DDFA-V2 and SynergyNet) emerged as the optimal tier under both balanced and accuracy-priority deployment profiles.

By eliminating the need for expensive physical hardware and complex, error-prone calibration protocols, the proposed digital twin framework offers a highly accessible, standardized, and mathematically rigorous methodology for stress-testing, vulnerability discovery, and reliability assessment of future human-centric perception and smart interactive systems.

7.2. Future Work

To address the identified limitations, several promising avenues of research will be pursued in future work:

First, expanding demographic diversity and environmental fidelity. To enhance benchmark diversity and mitigate potential demographic biases in facial perception, we plan to incorporate a larger, highly diverse collection of digital human identities spanning varied facial geometries, skin tones, age groups, and expressions. We will also introduce challenging physical environmental phenomena—such as dynamic illumination, motion blur, rolling-shutter distortion, simulated sensor thermal noise, and domain randomization—to further bridge the gap between simulation and real-world physical deployments.

Second, expanding algorithmic coverage and reality gap validation. The benchmark will be expanded to evaluate recent Transformer-based facial perception architectures, multi-modal RGB-depth fusion networks, and spatial perception models. Furthermore, to systematically analyze the transition from simulation to reality, future efforts will focus on validating simulation-to-real-world consistency by integrating highly controlled physical camera experiments with our virtual benchmarking pipeline.

Third, elevating statistical diagnostics. To elevate our statistical diagnostics, we will investigate the application of Linear Mixed-Effects Models (LMMs). This statistical approach will enable us to quantitatively decouple and analyze the complex, non-linear interaction effects between head orientation angles, spatial translational offsets, and algorithmic architecture, thereby providing a more mathematically rigorous model of tracking reliability.

Author Contributions: Conceptualization, K.T. and C.C.; methodology, K.T. and C.C.; software, K.T., M.S.A. and Z.L.; validation, K.T., M.S.A. and Z.L.; formal analysis, K.T. and Z.L.; investigation,

K.T.; resources, C.C., K.M. and K.C.C.; data curation, K.T., M.S.A. and Z.L.; writing—original draft preparation, K.T.; writing—review and editing, C.C., K.M., K.C.C., M.S.A. and Z.L.; visualization, K.T., S.A. and M.S.A.; supervision, C.C., K.M. and K.C.C.; project administration, C.C.; funding acquisition, C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Jilin Province Foreign Expert Program (2026), grant number L202601, under the project entitled “Application of Sim-to-Real Digital Twin Technology for 3D Eye Perception in Human–Machine Interaction for Intelligent Cockpits”. The APC was funded by the Jilin Province Foreign Expert Program (L202601).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Koesdwiady, A.; Soua, R.; Karray, F.; Kamel, M.S. Recent trends in driver safety monitoring systems: State of the art and challenges. *IEEE Trans. Veh. Technol.* **2016**, *66*, 4550–4563. [CrossRef]
2. Clay, V.; König, P.; Koenig, S. Eye tracking in virtual reality. *J. Eye Mov. Res.* **2019**, *12*, 3. [CrossRef]
3. Adhanom, I.B.; MacNeilage, P.; Folmer, E. Eye tracking in virtual reality: A broad review of applications and challenges. *Virtual Real.* **2023**, *27*, 1481–1505. [CrossRef] [PubMed]
4. Wang, K.; Ji, Q. Real time eye gaze tracking with 3d deformable eye-face model. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1003–1011.
5. King, D.E. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
6. Guo, J.; Zhu, X.; Yang, Y.; Yang, F.; Lei, Z.; Li, S.Z. Towards fast, accurate, and stable 3d dense face alignment. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 152–168.
7. Wu, C.Y.; Duan, Q.; Dai, Z.; Zhao, J.; Ruan, H.; Wei, M. Synergy-constrained 3d dense face alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 3721–3730.
8. Albiero, V.; Chen, X.; Yin, X.; Pang, G.; Hassner, T. img2pose: Face alignment and detection via 6dof pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 7617–7627.
9. Hempel, T.; Abdelrahman, A.A.; Al-Hamadi, A. 6d repnet: 6d head pose estimation in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3146–3154.
10. Wu, W.; Peng, H.; Yu, S. YuNet: A tiny millisecond-level face detector. *Mach. Intell. Res.* **2023**, *20*, 656–665. [CrossRef]
11. Krafka, K.; Ananya, A.; Wang, L.; Chang, A.; Shen, H.; Vedaldi, A.; Torralba, A. Eye tracking for everyone. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2176–2186.
12. Zhang, X.; Park, S.; Beeler, T.; Bradley, D.; Tang, S.; Hilliges, O. ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In Proceedings of the European Conference on Computer Vision (ECCV), Online, 23–28 August 2020.
13. Schuetz, I.; Fiehler, K. Eye tracking in virtual reality: Vive pro eye spatial accuracy, precision, and calibration reliability. *J. Eye Mov. Res.* **2022**, *15*, 15. [CrossRef]
14. Keselman, L.; Iselin Woodfill, J.; Grunnet-Jepsen, A.; Bhowmik, A. Intel realsense stereoscopic depth cameras. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017.
15. Wood, E.; Baltrušaitis, T.; Hewitt, C.; Dziadzio, S.; Cashman, T.J.; Shotton, J. Fake it till you make it: Face analysis in the wild using synthetic data alone. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 3681–3691.
16. Kortylewski, A.; Egger, B.; Schneider, A.; Gerig, T.; Morel-Forster, A.; Vetter, T. Analyzing and reducing the damage of dataset bias to face recognition with synthetic data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019.
17. Tremblay, J.; Prakash, A.; Acuna, D.; Brophy, M.; Jampani, V.; Anil, C.; To, T.; Cameracci, E.; Boochoon, S.; Birchfield, S. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 969–977.
18. Wood, E.; Baltrušaitis, T.; Zhang, X.; Sugano, Y.; Robinson, P.; Bulling, A. UnityEyes: Post-production-free 3D eye gaze estimation with a computer graphics pipeline. In Proceedings of the 9th ACM International Conference on Out-of-the-Box Interactive Technologies, Brisbane, Australia, 4–8 June 2016.

19. Wood, E.; Baltrušaitis, T.; Morency, L.P.; Robinson, P.; Bulling, A. Learning an appearance-based gaze estimator from one million synthesised images. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
20. Smith, A.D.; Muthumanickam, B.; Feng, Y.; Ding, W.; Hauser, K. UnityEyes 2: Open source synthetic eye generation for camera-based eye tracking with machine learning. In Proceedings of the 2025 Symposium on Eye Tracking Research and Applications (ETRA), Tokyo, Japan, 26–29 May 2025.
21. Byrne, S.A.; Niehorster, D.C.; Maquiling, V.; Nyström, M.; Kasneci, E. Digitally Prototype Your Eye Tracker: Simulating Hardware Performance using 3D Synthetic Data. *arXiv* **2025**, arXiv:2501.12345.
22. Egger, B.; Smith, W.A.; Tewari, A.; Wuhler, S.; Zollhoefer, M.; Beeler, T.; Bernard, F.; Bolkart, T.; Kortylewski, A.; Romdhani, S.; et al. 3D morphable face models—Past, present, and future. *ACM Trans. Graph. (TOG)* **2020**, *39*, 1–38. [[CrossRef](#)]
23. Zollhöfer, M.; Thies, J.; Garrido, P.; Bradley, D.; Beeler, T.; Pérez, P.; Theobalt, C.; Nießner, M.; Richardt, C. State of the art on monocular 3D face reconstruction, tracking, and rendering. *Comput. Graph. Forum* **2018**, *37*, 525–550. [[CrossRef](#)]
24. Lattas, A.; Sanyal, S.; Zafeiriou, S.; Ghosh, A. FitMe: Deep photorealistic 3D morphable model faces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 8677–8687.
25. Ghosh, A.; Fyfe, G.; Tunwattanapong, B.; Busch, J.; Yu, X.; Debevec, P. Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph. (TOG)* **2011**, *30*, 1–10. [[CrossRef](#)]
26. Remondino, F.; Fraser, C. Digital camera calibration methods: Considerations and comparisons. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2006**, *36*, 266–272.
27. Schönberger, J.L.; Zheng, E.; Frahm, J.M.; Pollefeys, M. Pixelwise view selection for unstructured multi-view stereo. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 501–518.
28. Schönberger, J.L.; Frahm, J.M. Structure-from-motion revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 4104–4113.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.