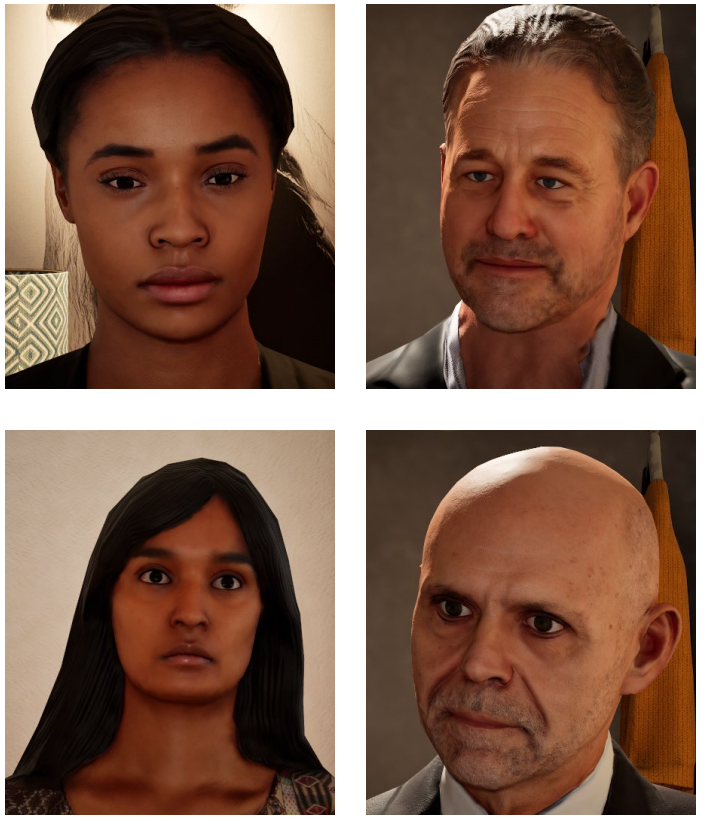
NeurIPS 2026 / Evaluations & Datasets / Paris

HumanStereo

A Benchmark for Metric Facial Depth Estimation and Its Evaluation



417,280 calibrated stereo pairs • task-aligned facial metrics

Alexandru I. Marinescu • Diana L. Borza • Adrian S. Dărăbant

Babeș-Bolyai University

Facial depth needs more than δ_1

Conventional stereo benchmarks emphasize large objects, not millimeter-scale facial geometry.

Why faces are difficult

1 Smooth, low-texture skin

Weak stereo correspondence on large facial areas

2 Small, protruding features

Eye corners, nose and lips are prone to occlusion

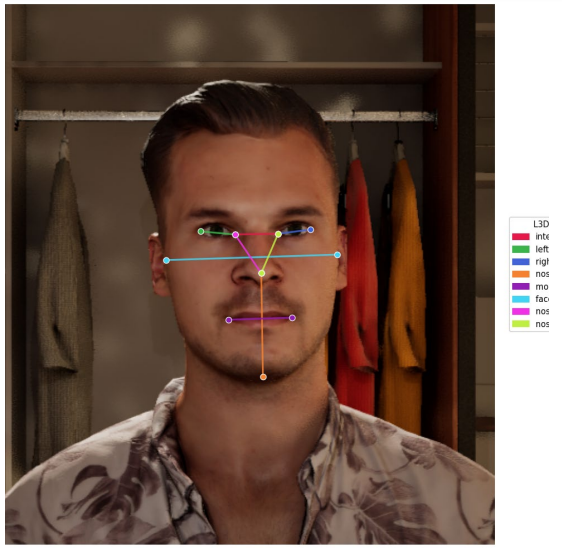
3 Absolute precision matters

IPD, eyewear fitting and AR/VR use metric distances

δ_1 = 1.25 permits 27.5 cm error at 110 cm!

Why task-grounded evaluation?

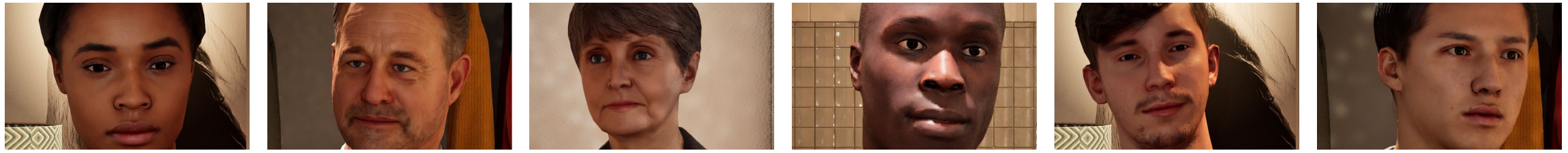
- Road and object benchmarks use different scales
- Relative scores hide mm-level depth errors
- Landmarks require accurate 3D distances



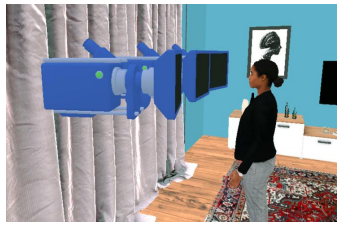
417,280 calibrated facial stereo pairs

Photorealistic avatars, diverse stereo geometry, dense annotations and exact calibration.

417,280	128	109	4
stereo sessions	3D human avatars	camera configurations	indoor environments



- Metric depth + surface normals
- Masks and 478 facial landmarks
- Intrinsics / extrinsics



Unreal Engine 5 controlled stereo rig

B: 5–15 cm • FOV: 10°–30°
Vergence: 0°, 1°, 2°
Distance: 20–200 cm • pose: ±10°

IR and L3DE expose what δ_1 hides

Complementary measures of absolute pixel accuracy and facial distances.

Absolute Inlier Ratio • IR(ε)

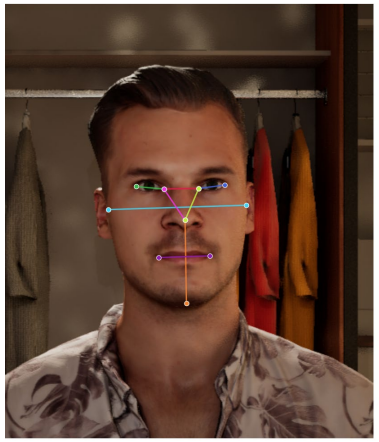
$$IR(\varepsilon) = \frac{1}{N}\sum_{i \in V} 1\left[\left|\widehat{d_i} - d_i\right| < \varepsilon\right]$$

ε ∈ {1, 2, 5, 10, 20} mm

- Percentage of pixels within an absolute metric tolerance
- 1 mm directly tests precision needed by fine facial tasks

Landmark 3D Distance Error • L3DE

$$L3DE = \frac{1}{|P|}\sum_{(i,j) \in P}\left|\left\|\widehat{p_i} - \widehat{p_j}\right\|_2 - \left\|p_i - p_j\right\|_2\right|$$



Eight anatomical landmark pairs, measured in millimeters.

Examples: interocular distance, eye width, mouth width, nose-to-chin.

Near-perfect δ_1 can conceal large errors

Performance on the humanoid evaluation region.

Method	δ_1 (%) ↑	MAE (mm) ↓	IR(1mm) ↑	L3DE (mm) ↓
BM	7.05	1663.11	0.86	377.49
SGBM3W	63.83	678.45	14.66	304.249
RAFT-Stereo	99.60	6.61	36.41	3.786
IGEV++	99.54	9.38	34.66	3.262
MASt3R	23.65	477.94	0.13	27.977
FoundationStereo	99.63	6.97	45.81	1.919
RAFT-Stereo ft	**99.64**	**4.99**	**48.32**	**1.840**
IGEV++ ft	99.58	6.89	38.41	2.927

Biggest takeaways

48.32%

best IR at 1 mm

1.840 mm

best landmark-distance error

RAFT-Stereo fine-tuned leads both. More than half of facial pixels still miss the 1 mm tolerance.

Region and stereo geometry change the story

Model errors depend strongly on where and how they are measured.

MAE (mm) by anatomical region

	Full	Humanoid	Face	Eye	Mouth
RAFT-Stereo ft	30.4	5.0	1.7	1.1	1.1
FoundationStereo	100.0	7.0	2.2	2.1	1.4
IGEV++	442.5	9.4	3.2	28.8	2.3

Full-scene averages hide fine facial-region behavior.

Vergence affects FoundationStereo

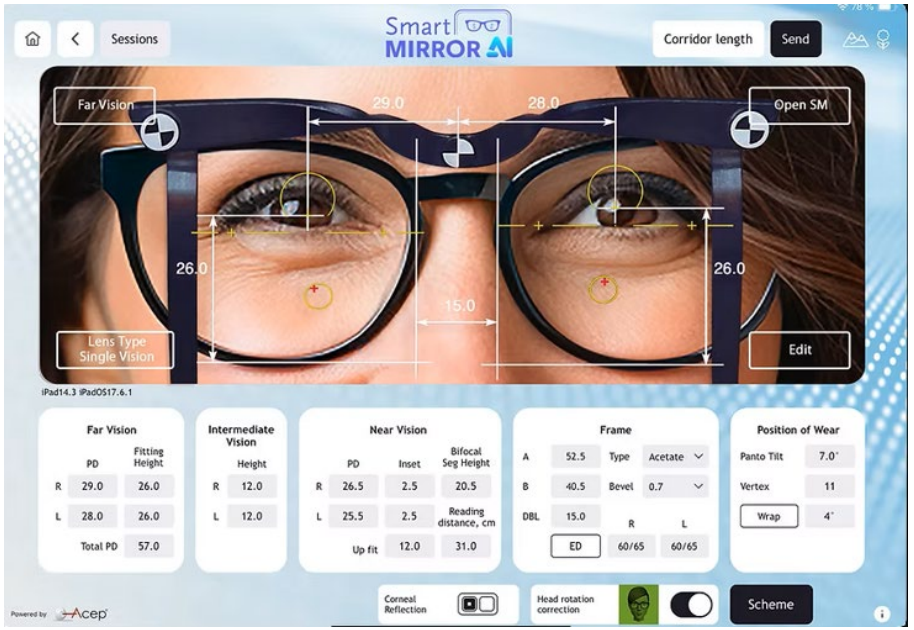
MAE increases as camera convergence grows:



Wide FOV and vergence can degrade performance even on the same subjects.

Real optometric IPD remains difficult

Ninety captures from four optometric devices.



Illustrative device UI (not a study participant)

Absolute IPD error • real captures

Method	IPD error (mm) ↓
RAFT-Stereo	7.06
RAFT-Stereo ft	5.71
IGEV++	2.96
IGEV++ ft	**1.94**

Best: 1.94 mm • still above sub-mm tolerance

What HumanStereo tells us

A benchmark, task-aligned evaluation, and a realistic measure of remaining challenges.

DATA	**417,280 calibrated sessions from 128 avatars.**
EVALUATION	**IR and L3DE reveal errors that δ_1 hides.**
GEOMETRY	**Camera setup and facial region change the ranking.**
TRANSFER	**1.94 mm real IPD error: not yet clinical-grade.**

Explore HumanStereo

huggingface.co/datasets/
asergiu/HumanStereo

Acknowledgements

This research is supported by the project Romanian Hub for Artificial Intelligence – HRIA, 2021–2027, MySMIS no. 351416, and project PN-IV-P7-7.1-PTE-2024-0105, contract no. 43PTE from 05/05/2025.

Thank you!