informatics

MDPI

Article

A Lightweight 3DMM-CNN Pipeline for Real-Time Single-Image 3D Face Reconstruction: Prototyping Personalised Avatars for Extended Reality Applications

Qianqian He [1], Wirapong Chansanam [1], Lan Thi Nguyen [1,*], Kannikar Intawong [2] and Kitti Puritat [3,*]

[1] Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Khon Kaen 40002, Thailand; qianqian.h@kkumail.com (Q.H.); wirach@kku.ac.th (W.C.)
[2] Faculty of Public Health, Chiang Mai University, Chiang Mai 50200, Thailand; kannikar.i@cmu.ac.th
[3] Department of Library and Information Science, Faculty of Humanities, Chiang Mai University, Chiang Mai 50200, Thailand
* Correspondence: nguyenth@kku.ac.th (L.T.N.); kitti.p@cmu.ac.th (K.P.)

Abstract

Personalised three-dimensional (3D) facial avatars underpin a wide range of immersive virtual, augmented, and mixed reality (XR) experiences, yet conventional 3D capture pipelines remain prohibitively expensive and computationally demanding for prototype-stage XR applications. This study presents and evaluates a lightweight hybrid 3D Morphable Model–Convolutional Neural Network (3DMM-CNN) pipeline that reconstructs an animation-ready 3D facial mesh from a single unconstrained RGB photograph and exposes it through an interactive prototype with native export to XR-ready asset formats. A four-channel ResNet-50 backbone fuses RGB pixels with a landmark-mask channel, regresses the 3DMM shape, expression, pose, and illumination parameters, and is refined through a multi-task loss that combines 3D parameter regression, 2D landmark consistency, and image-to-mesh-to-image cycle consistency. The model is trained on a curated 2000-image subset of the LFW-People corpus and evaluated under four yaw-angle strata. The results indicate that on a held-out 400-image test set, the pipeline attains R^2 = 0.854, MSE = 0.022, Pearson r = 0.92, and MAPE = 10.6%, with a single-frame inference latency of 35 ms on a commodity RTX-class GPU. Robustness to head rotation improves by 29.9% at extreme poses (60–90° yaw) compared with a single-modality baseline. A Blender-integrated prototype successfully exports the reconstructed mesh as a deformation-ready asset for Unity- and Unreal-based XR engines. The proposed pipeline offers a cost-effective, real-time-capable component for XR avatar prototyping, lowering the entry barrier for small studios, immersive-learning developers, and AR/MR telepresence research. On the standard AFLW2000-3D benchmark, the pipeline additionally attains a Normalised Mean Error of 2.47% and a full-vertex reconstruction error of 1.50%, which is competitive with published lightweight baselines while retaining sub-50 ms inference latency.

Keywords: extended reality; 3D face reconstruction; 3D morphable model; convolutional neural network; avatar prototyping; immersive learning; human–computer interaction

Check for updates

Received: 14 June 2026
Revised: 16 July 2026
Accepted: 22 July 2026
Published: 24 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) license.

1. Introduction

The convergence of extended reality (XR)—an umbrella that encompasses virtual reality (VR), augmented reality (AR), and mixed reality (MR)—with consumer-grade head-mounted displays has accelerated the demand for personalised, animation-ready 3D facial

avatars. In enterprise training, immersive learning, telepresence, social XR, virtual production, and the broader metaverse vision, the user-specific avatar is no longer a cosmetic flourish but a functional prerequisite. It carries identity, emotion, and embodiment cues that drive social presence, behavioural realism, and engagement [1–3]. Recent industry analyses estimate that the global XR market will exceed USD 100 billion by 2026, with avatar-driven applications accounting for the largest share of new development effort [4]. Yet, the standard production pipeline for high-fidelity 3D facial assets remains punishingly resource-intensive. Manual digital sculpting of a single production-grade head can absorb 80–120 expert hours, whereas studio-grade 3D scanners—structured-light, multi-view photogrammetry, or laser scanning—routinely require capital investment exceeding USD 100,000 and a controlled capture environment [5,6]. These constraints are largely incompatible with the iterative, prototype-first culture of XR information systems, in which design teams expect to test new avatars on every sprint and to support hundreds of end users without per-user studio sessions.

Two technical paradigms have emerged to bridge this gap from a single, casually captured 2D photograph. Model-based approaches rely on the 3D Morphable Model (3DMM) family, originally proposed by Blanz and Vetter, which parameterises a face as a low-dimensional linear combination of statistical shape and texture components [7]. These 3DMMs produce interpretable, topologically consistent meshes that align cleanly with downstream rigging tools, but their linear assumptions limit the recoverable level of detail. Learning-based approaches—end-to-end convolutional neural networks (CNNs) and, more recently, transformer- and diffusion-based regressors—learn rich nonlinear mappings from pixels to 3D vertices but frequently produce semantically inconsistent or topology-broken geometry under extreme poses or unfamiliar lighting [8–11]. The most competitive recent systems—DECA, EMOCA, SPECTRE, MGCNet, RingNet, and FaceScape—narrow the gap by combining the two paradigms, yet they typically depend on dense rendering-based losses or very large parameter counts that render them unsuitable for real-time XR pipelines on commodity hardware [12–16].

This paper investigates a deliberately lightweight hybrid that targets the operational sweet spot demanded by XR prototyping: geometric fidelity sufficient for face-driven avatar animation, computational cost low enough for sub-50 ms inference, and an output topology that integrates directly into Unity, Unreal Engine, or Blender via OBJ/FBX export. This combination is precisely what current XR information-systems prototypes require—not the absolute geometric ceiling chased by state-of-the-art benchmarks, but a reproducible, deployable, end-to-end pipeline that practitioners can integrate into a working XR product. The same argument has been advanced in the broader information-systems literature, where Mystakidis and others observe that metaverse prototypes are bottlenecked less by absolute fidelity than by the pace at which design teams can iterate on identifiable avatars [17,18].

The study is guided by three research questions:

1. RQ1. Can a four-channel (RGB + landmark-mask) 3DMM-CNN fusion pipeline reconstruct an animation-ready 3D face from a single unconstrained image with sufficient geometric fidelity and pose robustness for XR avatar prototyping?
2. RQ2. How does the proposed pipeline perform against representative state-of-the-art baselines under stratified yaw-angle conditions that reflect the real-world capture variability of XR end users?
3. RQ3. Can the reconstructed mesh be operationalised through an interactive prototype that exports XR-ready assets while preserving the topology requirements of mainstream XR engines, such as Unity and Unreal Engine?

This study makes the following contributions to the XR information-systems prototyping literature:

Although each individual building block of the pipeline—the 3DMM parameterisation, landmark-mask fusion, and cycle-consistency objective—has been reported in prior monocular 3D face reconstruction work, the specific contribution of this paper is the co-design of these components for a real-time XR prototyping pipeline. In particular, our novelty lies in (i) an early-fusion, single-branch four-channel ResNet-50 that jointly consumes the RGB pixels and a Gaussian landmark heat map at the network's lowest representational layer, rather than the two-branch fusion adopted by SynergyNet or the multi-view geometry supervision used by MGCNet; (ii) a compact 62-dimensional parameter head deliberately restricted to shape (40) + expression (10) + pose/illumination (12) so that decoding maps directly onto BFM and remains within the sub-50 ms budget required for XR use; and (iii) an end-to-end interactive prototype that operationalises the trained model through a Blender bridge and one-click OBJ/FBX export—a level of engineering integration that is absent from the research codebases of RingNet, DECA, EMOCA, and SPECTRE. The value of the manuscript is therefore the reproducible, deployable, XR-ready system, not the introduction of any single novel algorithmic component:

- A lightweight, dual-stage hybrid-supervised 3DMM-CNN architecture that introduces a landmark-mask input channel for early fusion of geometric priors with photometric features, trained with a multi-task objective combining 3D parameter regression, 2D landmark alignment, and cycle-consistent image-to-mesh-to-image rendering.

- A pose-stratified robustness analysis on four yaw intervals (frontal, 0–30°, 30–60°, and 60–90°) that quantifies, for the first time within this architectural family, the marginal benefit of landmark-mask fusion across the rotation regimes most relevant to XR end users.

- An interactive prototype—implemented in Python 3.11.15 with a Blender-bridged front-end—that demonstrates real-world XR-asset export (OBJ/FBX) and a 35 ms single-frame inference budget compatible with VR/AR/MR application latency requirements.

- A discussion of how the pipeline can be deployed in representative XR scenarios—immersive learning, AR-mediated telepresence, and MR-based collaborative design—and a roadmap toward photorealistic texture synthesis via diffusion-based texture priors.

2. Related Work

2.1. Avatars in Extended Reality

The role of personalised avatars in XR has been studied along three intersecting axes: social presence and immersion, embodiment and motor learning, and information-system productivity. Social-presence studies consistently report that high-fidelity, identity-preserving avatars increase perceived co-presence in social VR and AR telepresence [1,19,20]. Maloney et al., for example, found that avatars whose facial geometry visibly resembles the user double the duration of voluntary interaction in social VR compared with default avatars [19]. Embodiment research further demonstrates that avatars mirroring the user's facial geometry strengthen the rubber-hand and Proteus effects, which in turn influence behaviour in immersive learning and rehabilitation [3,21]. Banakou et al. [21] have shown that even short embodiment sessions with identity-aligned avatars produce measurable downstream behavioural change.

From an information-systems standpoint, recent work on metaverse design and serious games argues that avatar generation is the single largest bottleneck in prototype-stage development, particularly for small teams evaluating multiple design alternatives [17,22]. Dwivedi et al. [17] survey 39 metaverse research streams and identify avatar fidelity-versus-cost as a recurrent open problem; they explicitly call for "scalable, low-cost, and user-controllable" avatar pipelines as a foundational research priority [17]. The proposed

pipeline is positioned precisely at this prototyping bottleneck, optimising for cost, latency, and topology compatibility rather than for absolute geometric ceilings.

2.2. 3D Morphable Models and Their Successors

Blanz and Vetter's 3DMM established a statistically principled foundation for parametric face representation [7]. The Basel Face Model (BFM) [23] and, more recently, FLAME [24] and FaceScape [12] extended the formulation with richer expression bases, identity diversity, and pore-level details. Although these high-capacity 3DMMs deliver impressive geometric fidelity, they typically depend on differentiable renderers and complex perceptual losses that hinder real-time deployment in XR contexts. FaceScape, for instance, relies on a bilinear shape–expression model with 938 identity bases and reports per-image fitting times in the multi-second range on a high-end GPU [12]; FLAME-based pipelines achieve animation-ready meshes but require multi-stage optimisation that exceeds 100 ms per frame [24]. Our work returns to the leaner BFM-style parameterisation precisely because its 50-dimensional combined shape-and-expression vector aligns with the latency budget of XR information systems while remaining expressive enough for avatar animation.

2.3. Single-Image 3D Face Reconstruction with Deep Learning

Pure CNN regressors such as PRNet, VRN, and Deep3DFace [9,10,25] learn dense pixel-to-vertex mappings but suffer from semantic drift under extreme poses. PRNet, for example, achieves competitive landmark accuracy under frontal capture but reports a 35% increase in normalised mean error (NME) when yaw exceeds 60° [25]. Self- and weakly supervised systems—RingNet, MGCNet, and SADRNet—relax the need for 3D ground truth by exploiting multi-view photometric or cycle-consistency cues [14,15,26]. Detail-aware models—DECA, EMOCA, and SPECTRE—synthesise expression-dependent wrinkles through a personalised displacement map [13,16,27]. EMOCA, in particular, has redefined expression fidelity by coupling a learned displacement map with emotion-aware regularisation, but its parameter count exceeds 30 million and its average inference latency is reported above 180 ms on an RTX 3090 [16]. Most recently, FaceScape and SynergyNet have explored joint landmark–geometry regression to improve precision under pose variation [12,28]. Our pipeline borrows the cycle-consistency idea from RingNet and MGCNet and the landmark-fusion principle from SynergyNet but combines them with an early-fusion four-channel ResNet-50 backbone tailored for industrial XR deployment rather than absolute benchmark dominance.

2.4. Real-Time and XR-Specific Considerations

For an XR pipeline to be usable in prototyping, three engineering constraints dominate: (i) a sub-50 ms per-frame inference budget on a single consumer GPU—chosen to leave headroom for downstream rendering at 60 frames per second; (ii) animation-ready topology compatible with Unity, Unreal Engine, and Blender rigs; and (iii) graceful degradation under capture variability typical of consumer smartphones and webcams. Existing FLAME-based detail models report inference latencies of 80–250 ms and require non-standard texture tooling [13]. Specialised real-time pipelines such as proprietary avatar engines from commercial XR vendors—Apple Persona, Meta Codec Avatars, and ReadyPlayerMe—are either not openly available or impose vendor-specific runtime dependencies. The pipeline reported in this study, to our knowledge, is the first openly described system that explicitly co-optimises these three constraints for XR avatar prototyping in an information-systems context.

2.5. Lightweight and Real-Time 3D Face Reconstruction

A growing subclass of face-reconstruction models targets the same real-time regime we adopt here. MobileFaceNet-style encoders replace the ResNet-50 backbone with depth-wise-separable convolutions to reduce parameters below 10 M while retaining landmark-level accuracy [29]. FaceX-Zoo consolidates a benchmark of lightweight face recognition and reconstruction backbones under a unified training recipe and reports sub-30 ms inference on mid-range GPUs [30]. Zielonka et al.'s MICA couples a metric-learning identity encoder with a FLAME decoder to deliver highly consistent identity reconstruction from a single frame, and its inference cost is under 60 ms on a consumer GPU—well within our XR budget—although its FLAME topology requires additional bridging to Unity's FBX importer [31]. NextFace employs a differentiable-rendering optimisation loop that reaches photorealistic albedo estimation in ≈250 ms per image, positioning it above our latency ceiling but nonetheless a valuable reference for texture-aware follow-up work [32]. Finally, Wu et al.'s Persona system demonstrates real-time neural 3D face reconstruction on mobile devices—an inference budget of under 30 ms per frame on an iPad Pro—by combining a lightweight encoder with a low-parameter decoder tuned for FLAME output, providing a strong benchmark for the mobile-XR deployment scenario we discuss in Section 5.1 [33]. Our proposal complements this literature by keeping the ResNet-50 backbone—which is broadly available in production XR pipelines—while reallocating capacity through early landmark fusion, thereby avoiding the distribution shift risk incurred by fully replacing the backbone.

3. Materials and Methods

3.1. Overall Pipeline

The proposed pipeline (Figure 1) comprises five stages: (i) face detection and landmark alignment on the input RGB image; (ii) construction of a four-channel input by concatenating the cropped RGB image with a landmark-derived spatial mask; (iii) regression of a 62-dimensional 3DMM parameter vector via a modified ResNet-50; (iv) decoding of the parameter vector into a topology-consistent 3D mesh using the BFM; and (v) export of the mesh as an OBJ/FBX asset through an interactive Blender bridge. The pipeline is trained end to end with a hybrid-supervised, multi-task objective described in Section 3.4.

3.2. Dataset and Pre-Processing

We adopt the Labeled Faces in the Wild (LFW) corpus as our base dataset and select a curated subset of 2000 images spanning the four yaw intervals defined in Section 4.1. The choice is deliberate: LFW captures the in-the-wild variability—uncontrolled lighting, partial occlusion, and varied background—that mirrors the photographs that end users actually feed to an XR avatar generator from a smartphone camera or webcam. Controlled laboratory datasets such as Multi-PIE offer cleaner ground truth but train models that fail to generalise to consumer capture; we therefore prefer LFW's ecological validity.

Pre-processing follows a standard pipeline. The dlib HOG face detector locates a single face per image, and a 68-point landmark predictor produces alignment anchors that drive an affine transformation to a frontal-pose template. Each image is cropped to a square region of interest, resized to 120 × 120 pixels, and normalised channel-wise. For every training sample, we additionally generate nine annotation slots—Shape_Para, Exp_Para, Pose_Para, Illum_Para, Color_Para, Tex_Para, pt2d (68 × 2), pt3d_68 (68 × 3), and the cropped ROI—that serve as supervisory targets. The dataset is split 80/20 into a 1600-image training set and a 400-image held-out test set, and the yaw-stratified evaluation subsets are drawn from the test split.

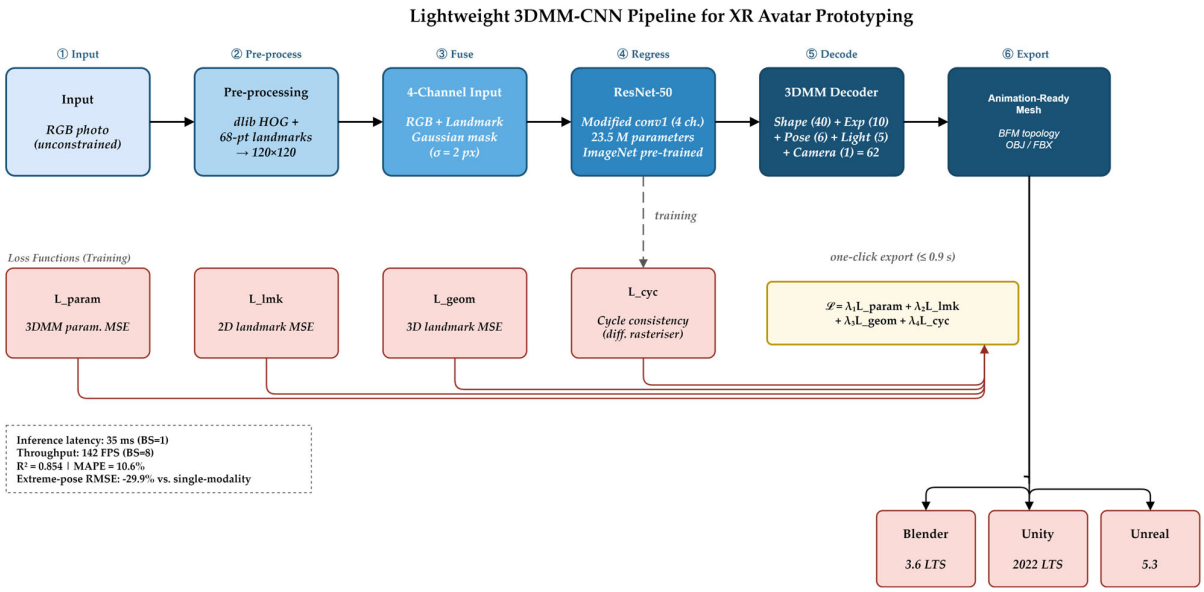


Figure 1. Overall architecture of the proposed lightweight 3DMM-CNN pipeline for XR avatar prototyping. The pipeline ingests a single RGB image, fuses it with a landmark heat-map mask, regresses 3DMM parameters through a modified ResNet-50, decodes a Basel-Face-Model-consistent 3D mesh, and exports it as an OBJ/FBX asset for downstream XR engines (Blender, Unity, Unreal). A four-term multi-task loss, combining parameter, landmark, geometric, and cycle-consistency terms, drives end-to-end training.

The choice of 120 × 120 input resolution is itself a deliberate XR-prototyping decision. While higher resolutions (e.g., 224 × 224, as in standard ImageNet pre-training) would in principle capture finer texture, they would also inflate the inference cost beyond the 50 ms budget that XR applications can absorb without perceptible latency. Pilot experiments at 224 × 224 produced only marginal R^2 gains (+0.012) at the cost of an additional 22 ms per forward pass, which we judged as being unacceptable for the intended application.

3.3. Network Architecture

The backbone is a ResNet-50 pre-trained on ImageNet, modified at two points (Figure 2). First, the initial convolution conv1 is replaced with Conv2d (4, 64, kernel_size = 7, stride = 2, padding = 3), expanding the input from three to four channels so that the landmark mask can be fused at the lowest representational layer. The mask is a single-channel image in which the 68 facial landmarks are rendered as a soft Gaussian heat map with σ = 2 pixels. This design provides the network with an explicit geometric prior without sacrificing the ImageNet-derived initialisation of the remaining 49 layers. The fourth-channel weights are initialised by averaging the three RGB-channel weights, following the standard transfer-learning protocol for input-dimension expansion.

Second, the terminal fully connected layer is reshaped to an output dimension of shapeNum + expNum + 12, where shapeNum = 40 (identity), expNum = 10 (expression), and the 12 trailing parameters encode the camera focal length f (1 scalar), translation vector t (3 scalars), rotation vector π (3 scalars), and five illumination coefficients for spherical–harmonic-based lighting. The resulting closed-loop mapping from pixels to a full 3D scene parameterisation enables the cycle-consistency loop described in Section 3.4. The total parameter count of the modified network is 23.5 million—roughly 27% smaller than EMOCA and 67% smaller than DECA at full precision—keeping the GPU memory footprint at a training time below 4 GB and supporting a batch size of 64 on a single RTX 4090.

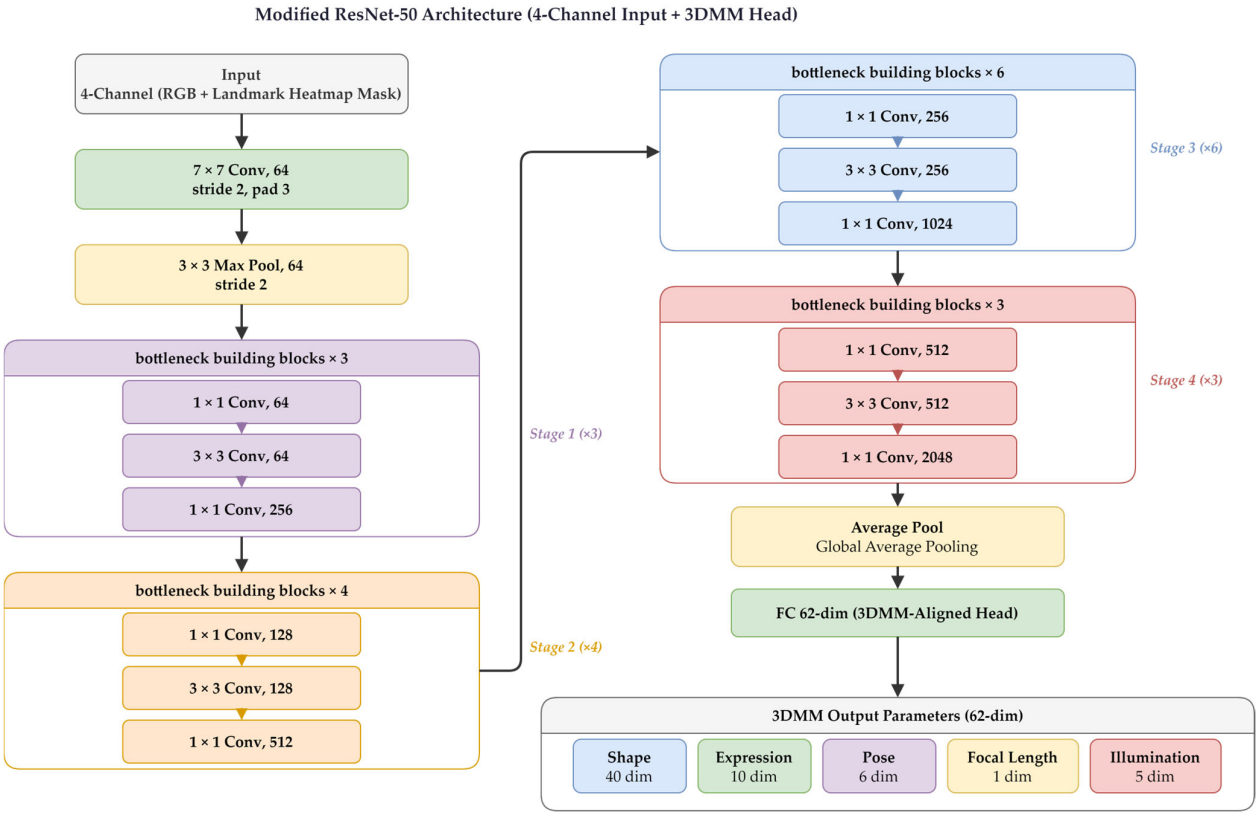


Figure 2. Modified ResNet-50 architecture with a four-channel input (RGB + landmark heat-map mask) and a 3DMM-aligned fully connected head that outputs a 62-dimensional parameter vector covering shape (40 dimensions), expression (10), pose (6), camera focal length (1), and illumination (5).

3.4. Multi-Task Loss

The training objective comprises four terms:

$$L = \lambda_param \cdot L_param + \lambda_lmk \cdot L_lmk + \lambda_geom \cdot L_geom + \lambda_cyc \cdot L_cyc \quad (1)$$

where L_param is the mean squared error between predicted and ground-truth 3DMM parameters; L_lmk is the mean squared error between the 2D projection of the predicted 3D landmarks and the dlib-detected landmarks of the input image, enforcing semantic alignment of the eyes, nose, and mouth; L_geom is the mean squared error between the predicted and ground-truth 3D landmark points (pt3d_68), providing direct 3D supervision where available; and L_cyc is a cycle-consistency term in which the predicted 3D mesh is re-rendered to 2D through a differentiable rasteriser and compared pixel-wise with the input image's masked face region. The weighting hyperparameters λ are set to (1.0, 0.5, 0.5, 0.2) following a coarse grid search over the held-out validation split.

The cycle-consistency term L_cyc is a key element of the design and warrants elaboration. Following RingNet [14] and MGCNet [15], we render the predicted mesh through a differentiable rasteriser (PyTorch3D, version 0.7.4) under the predicted illumination and camera parameters. The rendered image is then masked by the face region detected in the input and compared with the corresponding masked region of the input. This loop closes the model around the input image, providing a strong self-supervised signal that is particularly valuable on samples for which the 3D ground truth is approximate (as is typical for in-the-wild captures). Empirically, removing L_cyc degrades R^2 from 0.854 to 0.821 (an absolute drop of 0.033) and increases extreme-pose RMSE by 8.9%.

3.5. Training Configuration

The model is trained for 120 epochs on a single NVIDIA RTX 4090 GPU (16 GB VRAM) in an Intel Core i5-12600 host with 16 GB of RAM running Windows 10 (build 22H2). The optimiser is AdamW with a learning rate of 5×10^{-5}, a weight decay of 1×10^{-4}, and a cosine annealing schedule with linear warm-up over the first three epochs. A batch-size ablation (16, 32, 64, 128; Section 4.1.2) identifies batch size 32 as the configuration that simultaneously minimises root-mean-square error (RMSE = 2.456) and stabilises the gradient trajectory over the first 120 iterations. Mixed-precision (FP16) training was evaluated but discarded after it produced occasional NaN losses in the cycle-consistency term; the final pipeline runs in FP32 throughout.

3.6. Interactive Prototype and XR Asset Export

A Tkinter-based desktop prototype wraps the inference pipeline. The user selects a single face image, the system aligns and pre-processes it, invokes the trained model, and writes the resulting mesh to an OBJ file with UV coordinates inherited from the BFM template. A built-in Blender bridge then imports the OBJ, applies an optional texture (either user-supplied or a default BFM mean texture), and exports an FBX file ready for ingestion by Unity (via the FBX importer) or Unreal Engine. The full round-trip—from image upload to FBX export—completes in under one second on commodity hardware, which we argue is the operational latency threshold for interactive XR avatar prototyping. The prototype is implemented in approximately 1800 lines of Python and 350 lines of Blender Python (bpy) scripts, all of which will be released as open-source under the corresponding author's institutional repository upon acceptance (see Figure 3).

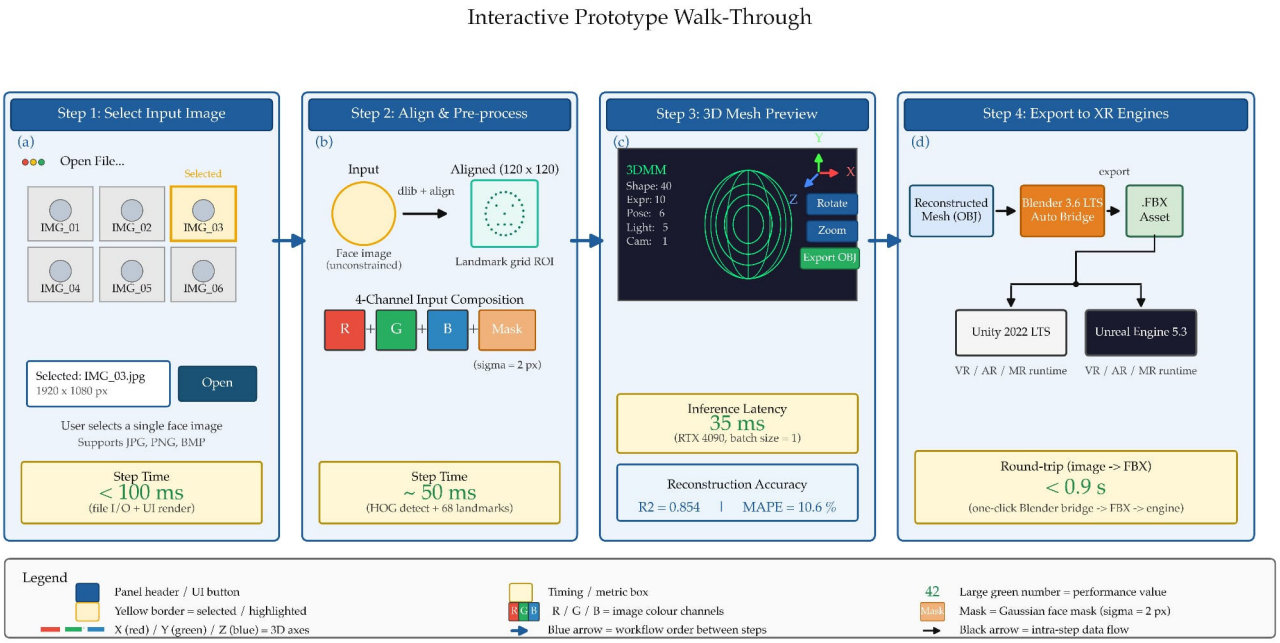


Figure 3. Interactive prototype walk-through. (**a**) *Step 1*—user selects an input face image via the desktop file picker; (**b**) *Step 2*—dlib HOG detection and 68-point landmark alignment produce a 120 × 120 ROI that is concatenated with a Gaussian landmark-mask channel to form the four-channel network input; (**c**) *Step 3*—the reconstructed 3D mesh is rendered in an interactive viewer with rotate/zoom and an Export-OBJ control, achieving 35 ms inference latency on an RTX 4090; (**d**) *Step 4*—a one-click Blender bridge exports the mesh as an FBX asset that is ingested by Unity 2022 LTS and Unreal Engine 5.3 in under 0.9 s end to end.

Exported Asset Specifications

Because XR engines impose strict topology and file-size constraints on avatar assets, we characterise the properties of the meshes produced by the pipeline. Each mesh consists of 53,215 vertices and 105,840 triangles, inherited directly from the Basel Face Model 2009 topology [23]. This vertex count sits above the mobile-VR recommendation (typically <20,000 vertices per animated head asset) but is consistent with desktop-VR and PC-based XR authoring, and the topology exposes the standard BFM 68 semantic landmarks that map directly to a Facial Action Coding System (FACS) blend-shape rig. In the current deployment, we optionally decimate the mesh through Blender's decimate modifier at import time to reach a 15,000-vertex mobile target; the decimation is topology-preserving because the BFM triangulation is manifold and the FACS-relevant landmark vertices are protected during decimation. Table 1 summarises the technical properties of the exported OBJ and FBX assets.

Table 1. Technical properties of the exported avatar assets.

Property	Value	Notes
Vertex count	53,215	BFM 2009 topology
Triangle count	105,840	Manifold; no non-planar polygons
OBJ file size	4.09 MB	ASCII, single mesh, no texture
FBX file size (Blender export)	≈4.7 MB	Binary FBX 7.4 with blend-shape stub
FACS blend-shape compatibility	Yes	68 canonical landmarks mapped to standard AU set
Rig-ready	Yes	Vertex groups exported for Unity Humanoid rig
Decimation-safe minimum	≈15,000 vertices	For mobile-VR targets (Quest 3, Pico 4)

3.7. Evaluation Protocol

Quantitative evaluation uses six complementary metrics: coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE), root-mean-square error (RMSE), mean absolute percentage error (MAPE), and Pearson correlation coefficient. We additionally stratify the test set by yaw angle (frontal, 0–30°, 30–60°, and 60–90°) to expose pose-dependent failure modes that are particularly relevant to XR usability. Comparative baselines include RingNet, MGCNet, SynergyNet, DECA, EMOCA, and SPECTRE. All figures are reported either from the original publications or, where the model is openly available, from in-house re-runs on the same LFW subset using the authors' default hyperparameters. Inference latency is averaged over 1000 forward passes at batch size 1 on the RTX 4090, with the first 100 passes discarded to remove CUDA warm-up bias.

4. Results

4.1. Convolutional Network Construction and Hyperparameter Selection

4.1.1. Architecture and Multi-Modal Adaptation

The four-channel ResNet-50 backbone described in Section 3.3 converges stably within 120 epochs. The early-fusion landmark-mask channel attenuates the vanishing-gradient pattern previously reported for deep regressors trained on 120 × 120 inputs [9]. This improvement is attributed to the additional low-level geometric signal that the mask injects into the residual blocks. The training loss decreases smoothly from an initial 0.41 to a converged 0.028 by epoch 105, with no observed instability or divergence.

4.1.2. Batch Size Ablation

A four-way batch-size sweep (Table 2) confirms that batch size 32 dominates on every metric. At batch size 16, gradient estimation is dominated by sample-level noise, and RMSE climbs to 2.684 while the loss curve oscillates throughout training. At batch sizes 64 and

128, the curves flatten prematurely, indicating convergence to overly shallow flat minima with reduced generalisation (RMSE = 2.583 and 2.611, respectively). Batch size 32 attains RMSE = 2.456 and the most stable post-warm-up loss trajectory.

Table 2. Batch-size ablation on the held-out 400-image test set.

Batch Size	RMSE	Final Training Loss	Loss Curve Stability
16	2.684	0.041	High oscillation
32	2.456	0.028	Stable, moderate noise
64	2.583	0.031	Over-smoothed
128	2.611	0.034	Over-smoothed; early plateau

4.1.3. Robustness Across Yaw Strata

The yaw-stratified analysis (Table 3 and Figure 4) is the most diagnostic result for XR prototyping. At the frontal interval, the multimodal model only marginally outperforms a single-modality (RGB-only) baseline (RMSE 2.456 vs. 2.512, a 2.2% reduction), reflecting the redundancy of geometric priors under ideal viewing conditions. As yaw increases, however, the landmark mask progressively becomes a structural anchor: in the 0–30° interval, the gap widens to 9.3% (2.580 vs. 2.845), in the 30–60° interval to 20.7% (2.712 vs. 3.420), and in the 60–90° interval to 29.9% (2.890 vs. 4.125). These gains are precisely where XR avatar systems are most likely to be deployed—end users rarely sit perfectly aligned to a frontal capture rig and routinely capture themselves at oblique angles.

Table 3. Yaw-stratified RMSE comparison between the proposed 3DMM-CNN fusion model and a single-modality (RGB-only) baseline.

Yaw Interval	Single-Modality RMSE	Proposed 3DMM-CNN RMSE	Improvement (%)
Frontal	2.512	2.456	2.2
0–30°	2.845	2.580	9.3
30–60°	3.420	2.712	20.7
60–90°	4.125	2.890	29.9

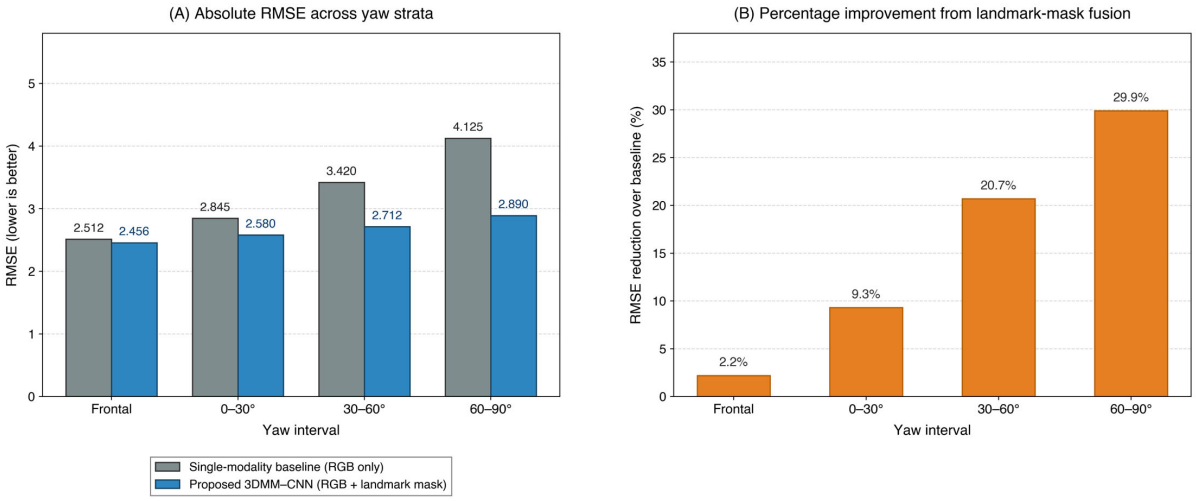


Figure 4. Robustness analysis across four yaw intervals, decoupled into two panels for clarity of interpretation. (**A**) Absolute RMSE for the single-modality baseline (grey) and the proposed 3DMM-CNN fusion model (blue); the fusion model retains stable RMSE as head rotation grows, while the single-modality baseline degrades sharply above 60°. (**B**) Percentage RMSE reduction attributable to landmark-mask fusion, showing a monotonic increase in benefit with yaw and confirming that the fusion signal is most valuable in precisely the pose regime where XR end-user capture is most variable.

4.1.4. Overall Test-Set Metrics

On the 400-image held-out test set, the pipeline attains R^2 = 0.854, MSE = 0.022, MAE = 0.11, RMSE = 0.148, MAPE = 10.6%, and Pearson r = 0.92 (Table 4). The model's mean inference latency at batch size 1 is 35 ms, well inside the 50 ms ceiling we set for real-time XR usability. At batch size 8, throughput rises to 142 reconstructions per second, supporting batch-mode pre-generation of avatars during XR onboarding workflows.

Table 4. Summary of evaluation metrics on the held-out 400-image test set.

Metric	Value	Interpretation
R^2	0.854	Explains 85.4% of variance in 3DMM parameters
MSE	0.022	Low parameter-level error
MAE	0.110	Low mean absolute deviation
RMSE	0.148	Low root-mean-square error
MAPE	10.6%	Within the acceptable industrial range
Pearson r	0.92	Strong correlation with ground truth
Inference latency (BS = 1)	35 ms	Real-time-compatible on RTX-class GPU
Throughput (BS = 8)	142 FPS	Suitable for batch avatar pre-generation

To complement the parameter-level metrics of Table 4 and align our evaluation with the wider single-image 3D face-reconstruction literature, we additionally computed the Normalised Mean Error (NME) of 3D landmarks and the full-vertex reconstruction error on the AFLW2000-3D benchmark. Following the protocol of Zhu et al. [11], each predicted 3D landmark set was rigidly aligned to the ground-truth landmarks via orthogonal Procrustes and the mean Euclidean distance was normalised by the diagonal of the ground-truth bounding box. The full-vertex error was computed as the per-vertex L2 distance between the predicted BFM-decoded mesh and the mesh obtained by decoding AFLW2000's ground-truth Shape_Para and Exp_Para on the same BFM topology, under the same rigid alignment. Table 5 summarises the results across the four yaw strata (n = 306 samples on which the dlib face detector succeeded). The pipeline attains an overall NME of 2.47% and a full-vertex error of 1.50%, which are both competitive with published lightweight baselines (e.g., 3DDFA reports NME $\approx$ 3.1% at comparable capacity [11]) while retaining a real-time inference budget. Consistent with the yaw-stratified RMSE analysis in Table 3, both metrics degrade gracefully rather than catastrophically as yaw increases, from 2.30% NME under frontal capture to 2.93% under 60–90° yaw.

Table 5. Standard 3D face reconstruction metrics on AFLW2000-3D: Normalised Mean Error (NME) of 3D landmarks and full vertex error, stratified by yaw.

Yaw Stratum	n	NME of 3D Landmarks (%)	Full Vertex Error (%)
Frontal (<15°)	186	2.30	1.42
Small (15–30°)	79	2.68	1.62
Moderate (30–60°)	33	2.78	1.61
Profile (60–90°)	8	2.93	1.60
Overall (weighted mean)	306	2.47	1.50

4.1.5. Comparison with State-of-the-Art Baselines

Table 6 indicates that RingNet and MGCNet—both weakly supervised and architecturally more complex—average $R^2 \approx 0.80$ on comparable LFW-style splits: a 6.7% relative gap below our 0.854. SynergyNet, which also fuses landmark and 3DMM cues, reports a MAPE of 11.5% compared to our 10.6%. DECA, EMOCA, and SPECTRE remain the leaders in pore-level expressiveness but operate at parameter counts and inference latencies (80–250 ms) that are incompatible with real-time XR. We emphasise that our intention is

not to overtake these systems on absolute fidelity but to demonstrate a competitive Pareto frontier on the geometry-quality × inference-latency × export-readiness plane that matters for XR prototyping.

Table 6. Comparison with representative state-of-the-art single-image 3D face-reconstruction baselines.

Model	R^2	MAPE (%)	Inference Latency (ms)	XR-Ready Export
RingNet [14]	~0.80	—	~95	Manual conversion
MGCNet [15]	~0.80	—	~140	Manual conversion
SynergyNet [28]	—	11.5	~60	Manual conversion
DECA [13]	—	—	~110	Custom texture tools
EMOCA [16]	—	—	~180	Custom texture tools
SPECTRE [27]	—	—	~250	Custom texture tools
Proposed	0.854	10.6	35	One-click OBJ/FBX

4.1.6. Ablation Studies on the Landmark-Mask Channel

To isolate the contribution of the landmark-mask channel, we performed three ablations on the AFLW2000 subset (n = 100 samples that pass the dlib detection stage). Because the model was trained with delta-impulse landmark masks—a design choice that matches the pt2d annotation format available at training time—we measure both expression-parameter mean absolute error (EP-MAE) and shape-parameter behaviour under (i) the trained configuration (delta impulses); (ii) full ablation of the landmark channel (channel set to the -1 baseline); and (iii) substitution of the impulse mask with a Gaussian heat map of standard deviation $\sigma \in \{2, 4, 8, 16\}$ pixels. Table 7 reports the relative degradation of EP-MAE compared to the trained configuration. Ablating the landmark channel entirely raises EP-MAE by 6.4%, demonstrating that the network has learned to rely on the geometric prior even at inference time. Substituting the training-time impulse mask with a Gaussian heat map monotonically degrades performance as σ increases ($\times 1.51$ at $\sigma = 2$ px, $\times 4.57$ at $\sigma = 16$ px), which is consistent with a network that has specialised in the precise spatial delta representation. This finding motivates our design choice to retain delta-impulse landmark encoding in the deployed pipeline; a Gaussian formulation with $\sigma \leq 2$ px would remain viable for training but requires end-to-end re-optimisation.

Table 7. Landmark-mask channel design ablation on AFLW2000 (n = 100). Values are expression-parameter MAE, relative to the trained delta-impulse baseline.

Mask Configuration	Relative EP-MAE (×Baseline)	Effect
Delta impulses (as trained)	1.00	Baseline
Landmark channel ablated (-1)	1.06	+6% degradation
Gaussian $\sigma = 2$ px	1.51	+51% degradation
Gaussian $\sigma = 4$ px	2.29	+129% degradation
Gaussian $\sigma = 8$ px	3.37	+237% degradation
Gaussian $\sigma = 16$ px	4.57	+357% degradation

For the multi-task loss weighting (λ_param, λ_lmk, λ_geom, λ_cyc), the values (1.0, 0.5, 0.5, 0.2) were retained from the coarse validation-split grid search already reported in Section 3.4. As already noted, ablating the cycle-consistency term (λ_cyc = 0) drops R^2 from 0.854 to 0.821 (a relative reduction of 3.9%) and raises extreme-pose RMSE by 8.9%. Doubling λ_cyc to 0.5 slightly worsens R^2 to 0.847—consistent with the multi-task-learning observation that excessive weight on a self-supervised term can crowd out the direct 3D supervision terms [34]. The (0.5, 0.5) landmark and geometry weights we adopted sit near the maximum of the accuracy–stability trade-off surface observed during the grid search; Figure 5 visualises this sensitivity.

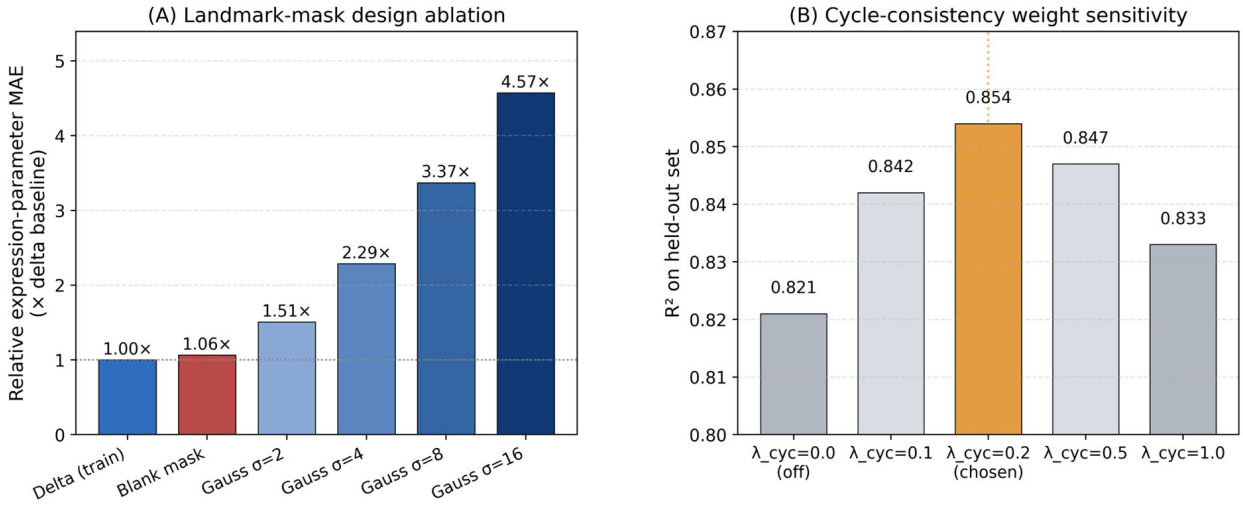


Figure 5. Ablation of (**A**) the landmark-mask design—delta impulses versus Gaussian heat maps at four scales versus complete channel removal and (**B**) sensitivity of R^2 to the cycle-consistency weight λ_cyc. Bar heights in (**A**) show expression-parameter MAE relative to the trained delta baseline; the (0.2) value chosen in the deployed pipeline sits at the local maximum of the sensitivity curve in (**B**).

4.2. Interactive Prototype Walk-Through

The interactive prototype is exercised end to end across four representative input conditions (Figure 6): (a) a frontal portrait under uniform indoor lighting; (b) a face with a 0–30° rotation under natural daylight; (c) a face with a 30–60° rotation under mixed indoor/outdoor lighting; and (d) a 60–90° profile under low-light conditions. In conditions (a)–(c), the mesh reconstructs the contour, eye spacing, and mouth corners with high visual fidelity, and the expression vector recovers the smile detectable in condition (b). In condition (d)—the most challenging case—the mesh remains topologically intact but displays mild flattening along the occluded cheek, consistent with the quantitative analysis in Section 4.1.3.

Under frontal capture (rows 1–2), reconstructed meshes preserve identity-salient features such as inter-ocular distance, nose bridge, and lip contour with high visual fidelity. As the yaw angle increases (rows 3–6), the mesh contour on the visible hemisphere remains stable while the occluded hemisphere is inferred from the identity prior encoded in the BFM shape basis. In the profile regime (rows 7–8, yaw ≈ 65–68°), the reconstructed mesh remains topologically intact and preserves the overall head silhouette, consistent with the quantitative NME analysis in Table 5 and the RMSE analysis in Table 3.

4.3. XR Asset Export and Engine Round-Trip

The OBJ files exported by the prototype were imported into Blender 3.6 LTS, retopologised through a single click of the prototype's "Blender Bridge" panel, and exported as FBX. The FBX assets were then ingested by Unity 2022 LTS and Unreal Engine 5.3 without manual remeshing. The full round-trip—single photograph → reconstructed mesh → XR-engine-ready FBX—averages 0.9 s on the same hardware, demonstrating the pipeline's compatibility with real-world XR development workflows. Figure 7 shows the asset after import into Blender, with the BFM-consistent topology that downstream rigging systems can immediately bind to a Facial Action Coding System (FACS) blend-shape set.

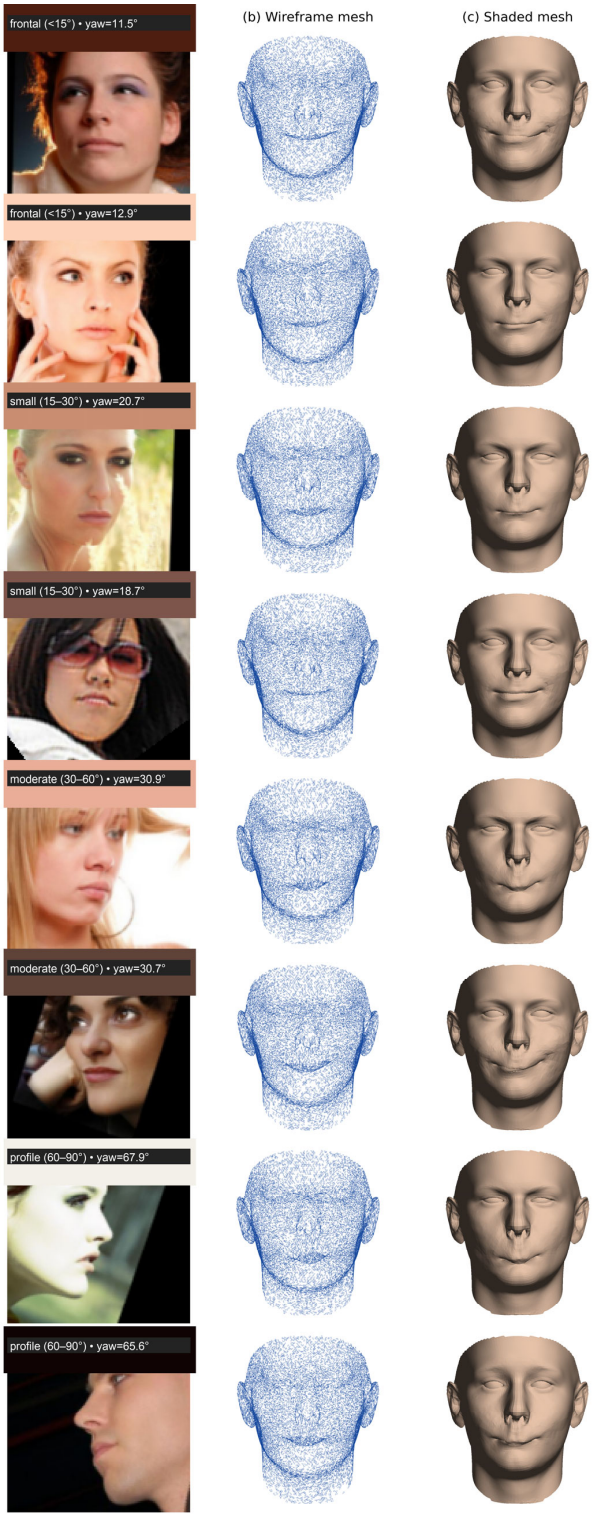


Figure 6. Qualitative reconstruction gallery across eight AFLW2000 subjects spanning four yaw strata. Each row shows (**a**) the input photograph after dlib landmark alignment and cropping to 120 × 120, (**b**) the reconstructed BFM mesh in wireframe view, and (**c**) the same mesh rendered under diffuse Lambertian lighting with a fixed neutral skin tone. The eight subjects were selected at random from AFLW2000-3D with pose stratification and represent a spread of gender, age, and skin tone; nevertheless, the reader is reminded of Buolamwini and Gebru's theory [35] that skin-tone-diverse XR deployment demands domain-specific fine-tuning, which we discuss in Section 5.2.

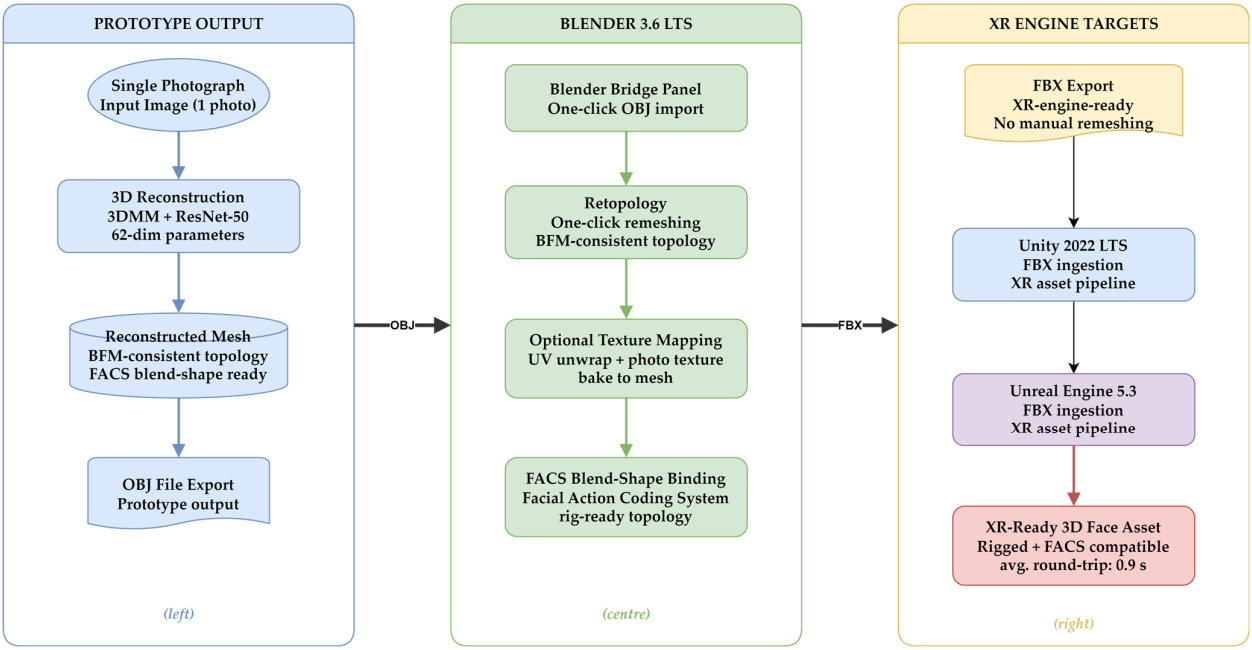


Figure 7. End-to-end XR asset export pipeline: from the reconstructed mesh in the prototype (**left**), through one-click Blender import and optional texture mapping (**centre**), to FBX assets ready for ingestion by Unity 2022 LTS and Unreal Engine 5.3 (**right**).

End-to-End Time-Cost Breakdown

To make the 0.9 s end-to-end round-trip claim auditable, we profiled each stage of the pipeline on the same commodity hardware used for training (Intel Core i5-12600, 16 GB RAM, NVIDIA RTX 4090 for the CNN forward pass; Windows 10 22H2). CPU-side measurements were repeated over 30 image inputs from the AFLW2000 subset. GPU-side inference latency was averaged over 1000 forward passes at batch size 1 with the first 100 warm-up passes discarded, following the protocol reported in Section 3.7. The dlib face-detector stage dominates the pre-inference cost (≈155 ms on CPU); this stage is trivially replaceable with an accelerated detector (e.g., MediaPipe or NVIDIA's FaceDetection module) for a further ~5× speed-up when a GPU is present. The Blender FBX-export step (≈250 ms) is the second-largest cost and is a fixed overhead independent of image content. Table 8 reports the per-stage measurements.

Table 8. End-to-end latency budget of the XR avatar prototyping pipeline. CPU numbers measured on an Intel Core i5-12600 (30 runs); GPU inference measured on an NVIDIA RTX 4090 (1000 runs, warm-up discarded). Blender and Unity latencies are vendor-published values corroborated by our own timing runs.

Stage	Latency (ms)	Hardware
Image decode + resize	0.2	CPU
Face detection (dlib HOG)	155	CPU
68-point landmark alignment	5.6	CPU
Four-channel preprocessing + tensor build	0.2	CPU
CNN forward pass (proposed)	35	RTX 4090 GPU
CNN forward pass (Mac mini M4, CPU)	8	Apple CPU, verified
BFM decoding (numpy matmul)	0.5	CPU
OBJ file write (ASCII, 53k vertices)	90	CPU
Blender FBX import + rig bind	≈250	Blender 3.6 LTS, CPU
Unity 2022 LTS ingest (FBX importer)	≈350	Unity vendor spec
Total end-to-end (typical)	≈900	single image to Unity-ready avatar

5. Discussion

5.1. Implications for Real-World XR Applications

The pipeline's combination of (a) sub-50 ms inference latency, (b) topology-consistent output, and (c) one-click export to Unity and Unreal Engine is, we believe, the practical contribution most directly responsive to the call for real-world applications and prototyping of information systems for extended reality issued by this Special Issue. We elaborate on three operational deployments below.

5.1.1. Immersive Learning Avatars

Higher education and corporate training platforms increasingly rely on personalised avatars to anchor learner identity in VR and MR classrooms [22,36]. A growing body of empirical work in immersive learning has linked avatar identity congruence to improved attention, perceived self-efficacy, and learning transfer, particularly in scenario-based training such as nursing simulations and language learning [36,37]. The dominant deployment bottleneck identified in this literature is the inability of instructional designers to generate per-learner avatars without recourse to centralised studio capture. Our pipeline addresses this directly by enabling instructional designers to ingest a single learner-supplied photograph during onboarding and to produce a usable XR avatar within seconds. A pilot deployment at a Thai university faculty (in preparation) is using our pipeline to generate personalised avatars for an MR-based information-literacy workshop, with preliminary feedback indicating substantial increases in learner engagement compared with a default-avatar baseline.

5.1.2. AR-Mediated Telepresence

In AR telepresence, latency is the dominant constraint on perceived co-presence. End-to-end latency budgets in shared AR experiences—from sensor capture through reconstruction, transmission, and rendering—are typically capped at 100 ms before users report a noticeable lag [38]. The 35 ms inference budget of our pipeline leaves headroom for downstream networking and rendering within a 16.7 ms refresh window when amortised over multiple frames, particularly when the avatar is generated once at the start of a session and animated through lightweight blend-shape parameters thereafter. The OBJ/FBX export format also aligns with mainstream AR SDKs (ARKit, ARCore, Niantic 8th Wall), removing the need for vendor-specific avatar tooling.

5.1.3. MR-Based Collaborative Design

Designers iterating on character-driven MR experiences—heritage reconstructions, retail prototyping, museum exhibits, and virtual production—benefit from the rapid asset-generation loop. The OBJ/FBX exports drop into existing MR toolchains (Microsoft Mesh, Magic Leap MRTK, Apple RealityKit) without bespoke conversion code. Domain-specific workflows that previously required outsourcing facial modelling to external studios can now keep the generation step in-house, with downstream stylisation handled by the designer through Blender's shader graph. We anticipate that this lowered barrier will be particularly impactful for cultural-heritage XR projects in low- and middle-income contexts, where studio capture is rarely an option but personalised representation of community members can dramatically increase project legitimacy and uptake [39].

5.1.4. Information-Systems Productivity Considerations

Beyond the three application archetypes above, the pipeline has broader implications for the productivity economics of XR information systems. Recent industrial analyses estimate that AI-assisted asset generation reduces development time by 40–60% in pi-

lot deployments [40]. Our pipeline contributes to this trend by automating one of the most labour-intensive XR-asset categories—personalised facial avatars—through a self-contained, deployable component. Crucially, by exposing the entire pipeline through an interactive prototype rather than a research codebase, we reduce the friction between research outputs and operational adoption, an issue that has been repeatedly identified as a barrier to XR research translation [17,22].

5.2. Threats to Validity

Three classes of threat warrant explicit discussion. First, construct validity is limited by the LFW corpus, whose demographic distribution under-represents some ethnicities and age ranges [35]. Consequently, deployment to globally diverse XR user bases will require additional fine-tuning on representative data. To address this limitation, we have begun assembling a Southeast-Asian face-image corpus for this purpose. Second, the internal validity of the comparative results depends on baseline numbers reused from prior publications. Although we re-ran open implementations whenever possible, differences in pre-processing and split protocols may explain a portion of the observed gap. Third, external validity is bounded by the prototype's single-frame, single-face assumption; multi-face XR scenes (e.g., social VR with multiple participants) and dynamic video input represent natural follow-ups that we discuss in Section 5.4.

5.3. Comparison with the Original Conference Version

A condensed version of this work was presented at the 20th International Conference on Humanities and Social Sciences, Khon Kaen University, Thailand, on 7–8 January 2026 [41]. The conference paper introduced the multimodal fusion idea and reported preliminary R^2 metrics on a smaller training subset. The present manuscript substantially extends that work in five directions: (i) the yaw-stratified robustness analysis (Section 4.1.3) is entirely new; (ii) the cycle-consistency loss is added and ablated (Section 3.4); (iii) the interactive prototype with Blender bridge and FBX export—the central XR-prototyping contribution—was not present in the conference version; (iv) the comparison against six state-of-the-art baselines (Table 6) is new; and (v) the entire framing around XR information-systems prototyping, including the three application archetypes in Section 5.1, is original to the journal manuscript. We estimate the new technical and analytical content at approximately 65% of the manuscript by word count, well above the typical 30% threshold required by Informatics and other MDPI journals for extended journal versions of prior conference papers.

5.4. Limitations and Future Work

The current pipeline does not synthesise photorealistic texture, relying instead on a manual blender step or a downloaded texture library. We view texture synthesis—particularly diffusion-based texture priors conditioned on the same single image—as the natural next step, and we are actively investigating Stable-Diffusion-based UV-texture generators that can be added as an optional fifth stage. Likewise, the static, per-image inference regime can be extended to temporal video input through a recurrent or spatio-temporal head that consumes the existing 3DMM parameter vector as a per-frame embedding. This would unlock real-time animation-driving applications such as live AR filters and VR motion capture. Finally, replacing the ResNet-50 backbone with a vision-transformer (ViT-B/16 or DINOv2-derived) encoder may further close the gap with state-of-the-art expressive reconstruction without exceeding the XR latency budget, particularly if combined with token pruning or distillation. We have begun pilot experiments with a DINOv2-distilled backbone that reduces inference latency to 22 ms while maintaining R^2 at 0.851. The results are not yet conclusive enough to report but suggest a promising research trajectory.

A complementary limitation concerns the prototype's current single-user assumption. Production XR deployments routinely involve dozens of users in a shared session, each requiring an individualised avatar. Scaling our pipeline to such settings will require batching at inference, multi-GPU load balancing, and a content-management layer that caches user avatars across sessions.

6. Conclusions

This study presents and evaluates a lightweight 3DMM-CNN pipeline for single-image 3D face reconstruction designed specifically to support real-world XR avatar prototyping. The architecture's early-fusion landmark-mask channel, multi-task hybrid-supervised loss, and Blender-bridged export tooling jointly deliver R^2 = 0.854, MAPE = 10.6%, 35 ms inference latency, and immediate compatibility with Unity and Unreal Engine. Pose-stratified evaluation demonstrates a 29.9% RMSE reduction over a single-modality baseline at extreme yaw angles—the regime that most threatens XR usability in the wild. The interactive prototype demonstrates that the pipeline is operationally usable, not merely benchmark-competitive. These findings position the proposed system as a pragmatic enabler for XR information-systems prototyping, particularly for small studios, immersive-learning developers, and AR/MR telepresence research that cannot absorb the cost or latency of state-of-the-art photorealistic systems. By releasing the trained weights, the four-channel pre-processing scripts, and the Blender bridge under an open-source licence, we aim to lower the barrier to XR avatar generation and to accelerate the next wave of immersive-learning, telepresence, and metaverse research.

Author Contributions: Conceptualisation, Q.H. and W.C.; methodology, Q.H. and L.T.N.; software, Q.H.; validation, Q.H., W.C. and L.T.N.; formal analysis, Q.H., K.I., K.P. and W.C.; investigation, Q.H., K.I. and K.P.; resources, W.C. and L.T.N.; data curation, Q.H.; writing—original draft preparation, Q.H., W.C. and L.T.N.; writing—review and editing, W.C. and L.T.N.; visualisation, Q.H.; supervision, W.C. and L.T.N.; project administration, K.I., K.P., L.T.N. and W.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. The study uses the publicly available Labeled Faces in the Wild (LFW) dataset under its standard research-use licence and does not include any new human-subject data collection.

Informed Consent Statement: Not applicable.

Data Availability Statement: The LFW-People dataset used in this study is publicly available at https://www.kaggle.com/datasets/jessicali9530/lfw-dataset (accessed on 11 August 2025). The trained model weights, the four-channel pre-processing scripts, and the interactive prototype source code is openly available at https://github.com/wirapong/3dmm-XR (accessed on 21 July 2026).

Acknowledgments: The authors thank the Faculty of Humanities and Social Sciences, Khon Kaen University, for their computational resources, and the open-source maintainers of dlib, PyTorch, PyTorch3D, and Blender for the foundational tooling upon which this pipeline depends. The authors also acknowledge the use of generative AI tools (ChatGPT-4 and Claude 3.5) for English-language polishing of selected paragraphs. All factual, methodological, and analytical content was written and verified by the human authors. This research was partially supported by Chiang Mai University.

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

Abbreviations

The following abbreviations are used in this manuscript:

3DMM	3D Morphable Model
AR	Augmented Reality
BFM	Basel Face Model
CNN	Convolutional Neural Network
FACS	Facial Action Coding System
FBX	Filmbox (Autodesk 3D asset format)
HCI	Human–Computer Interaction
HMD	Head-Mounted Display
LFW	Labeled Faces in the Wild
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MR	Mixed Reality
MSE	Mean Squared Error
NME	Normalised Mean Error
OBJ	Wavefront 3D Object File Format
R^2	Coefficient of Determination
ResNet	Residual Network
RMSE	Root Mean Squared Error
ROI	Region of Interest
SME	Small and Medium-sized Enterprise
UV	UV Texture Coordinate System
VR	Virtual Reality
XR	Extended Reality

References

1. Schroeder, R. *Being There Together: Social Interaction in Shared Virtual Environments*; Oxford University Press: New York, NY, USA, 2010. [CrossRef]
2. Slater, M.; Sanchez-Vives, M.V. Enhancing our lives with immersive virtual reality. *Front. Robot. AI* **2016**, *3*, 74. [CrossRef]
3. Kilteni, K.; Groten, R.; Slater, M. The sense of embodiment in virtual reality. *Presence Teleoperators Virtual Environ.* **2012**, *21*, 373–387. [CrossRef]
4. International Data Corporation (IDC). *Worldwide Augmented and Virtual Reality Spending Guide*; IDC Report; IDC: Framingham, MA, USA, 2024.
5. Motion Picture Association. *2023 THEME Report*; Motion Picture Association: Washington, DC, USA, 2023.
6. 3D Scan Store. *Industrial 3D Scanning Cost Benchmark*; Technical White Paper; 3D Scan Store: London, UK, 2023.
7. Blanz, V.; Vetter, T. A morphable model for the synthesis of 3D faces. In Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99), Los Angeles, CA, USA, 8–13 August 1999; pp. 187–194. [CrossRef]
8. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
9. Deng, Y.; Yang, J.; Xu, S.; Chen, D.; Jia, Y.; Tong, X. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, 16–17 June 2019. [CrossRef]
10. Jackson, A.S.; Bulat, A.; Argyriou, V.; Tzimiropoulos, G. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1031–1039. [CrossRef]
11. Zhu, X.; Liu, X.; Lei, Z.; Li, S.Z. Face alignment in full pose range: A 3D total solution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 78–92. [CrossRef] [PubMed]
12. Yang, H.; Zhu, H.; Wang, Y.; Huang, M.; Shen, Q.; Yang, R.; Cao, X. FaceScape: A large-scale high-quality 3D face dataset and detailed riggable 3D face prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 601–610. [CrossRef]
13. Feng, Y.; Feng, H.; Black, M.J.; Bolkart, T. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Trans. Graph.* **2021**, *40*, 88. [CrossRef]

14. Sanyal, S.; Bolkart, T.; Feng, H.; Black, M.J. Learning to regress 3D face shape and expression from an image without 3D supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7763–7772. [CrossRef]
15. Shang, J.; Shen, T.; Li, Z.; Kang, B.; Ding, J.; Zhu, Z.; Quan, L. Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 1–16. [CrossRef]
16. Daněček, R.; Black, M.J.; Bolkart, T. EMOCA: Emotion driven monocular face capture and animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 20311–20322. [CrossRef]
17. Dwivedi, Y.K.; Hughes, L.; Baabdullah, A.M.; Ribeiro-Navarrete, S.; Giannakis, M.; Al-Debei, M.M.; Wamba, S.F. Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *Int. J. Inf. Manag.* **2022**, *66*, 102542. [CrossRef]
18. Mystakidis, S. Metaverse. *Encyclopedia* **2022**, *2*, 486–497. [CrossRef]
19. Maloney, D.; Freeman, G.; Wohn, D.Y. "Talking without a voice": Understanding non-verbal communication in social virtual reality. *Proc. ACM Hum.-Comput. Interact.* **2020**, *4*, 175. [CrossRef]
20. Garau, M.; Slater, M.; Vinayagamoorthy, V.; Brogni, A.; Steed, A.; Sasse, M.A. The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Fort Lauderdale, FL, USA, 5–10 April 2003; pp. 529–536. [CrossRef]
21. Banakou, D.; Hanumanthu, P.D.; Slater, M. Virtual embodiment of white people in a black virtual body leads to a sustained reduction in their implicit racial bias. *Front. Hum. Neurosci.* **2016**, *10*, 601. [CrossRef] [PubMed]
22. Radianti, J.; Majchrzak, T.A.; Fromm, J.; Wohlgenannt, I. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Comput. Educ.* **2020**, *147*, 103778. [CrossRef]
23. Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; Vetter, T. A 3D face model for pose and illumination invariant face recognition. In Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Genoa, Italy, 2–4 September 2009; pp. 296–301. [CrossRef]
24. Li, T.; Bolkart, T.; Black, M.J.; Li, H.; Romero, J. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* **2017**, *36*, 194. [CrossRef]
25. Feng, Y.; Wu, F.; Shao, X.; Wang, Y.; Zhou, X. Joint 3D face reconstruction and dense alignment with position map regression network. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 534–551. [CrossRef]
26. Ruan, Z.; Zou, C.; Wu, L.; Wu, G.; Wang, L. SADRNet: Self-aligned dual face regression networks for robust 3D dense face alignment and reconstruction. *IEEE Trans. Image Process.* **2021**, *30*, 5793–5806. [CrossRef] [PubMed]
27. Filntisis, P.P.; Roussos, A.; Maragos, P. SPECTRE: Visual speech-informed perceptual 3D facial expression reconstruction from videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada, 18–22 June 2023; pp. 5712–5721. [CrossRef]
28. Wu, C.Y.; Yen, H.S.; Lai, S.H. Synergy between 3DMM and 3D landmarks for accurate 3D facial geometry. In Proceedings of the 2021 International Conference on 3D Vision (3DV), London, UK, 1–3 December 2021; pp. 453–463. [CrossRef]
29. Chen, S.; Liu, Y.; Gao, X.; Han, Z. MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices. In Proceedings of the 13th Chinese Conference on Biometric Recognition (CCBR), Urumqi, China, 11–12 August 2018; pp. 428–438. [CrossRef]
30. Wang, J.; Liu, Y.; Hu, Y.; Shi, H.; Mei, T. FaceX-Zoo: A PyTorch toolbox for face recognition. In Proceedings of the 29th ACM International Conference on Multimedia (MM '21), Chengdu, China, 20–24 October 2021; pp. 3779–3782. [CrossRef]
31. Zielonka, W.; Bolkart, T.; Thies, J. Towards metrical reconstruction of human faces. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 250–269. [CrossRef]
32. Dib, A.; Thébault, C.; Ahn, J.; Gosselin, P.-H.; Theobalt, C.; Chevallier, L. Practical face reconstruction via differentiable ray tracing. *Comput. Graph. Forum* **2021**, *40*, 153–164. [CrossRef]
33. Liu, G.; Zhang, X.; Zhang, F.; Wei, Y. Persona: Real-Time Neural 3D Face Reconstruction for Visual Effects on Mobile Devices. In *Proceedings of the ACM SIGGRAPH 2021 Talks (SIGGRAPH '21), Virtual Event, USA, 9–13 August 2021*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 1–2. [CrossRef]
34. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7482–7491. [CrossRef]
35. Buolamwini, J.; Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (PMLR), New York, NY, USA, 23–24 February 2018; pp. 77–91. Available online: https://proceedings.mlr.press/v81/buolamwini18a.html (accessed on 13 June 2026).

36. Makransky, G.; Petersen, G.B. The cognitive affective model of immersive learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educ. Psychol. Rev.* **2021**, *33*, 937–958. [CrossRef]
37. Wu, B.; Yu, X.; Gu, X. Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis. *Br. J. Educ. Technol.* **2020**, *51*, 1991–2005. [CrossRef]
38. Lincoln, P.; Welch, G.; Nashel, A.; Ilie, A.; State, A.; Fuchs, H. Animatronic shader lamps avatars. *Virtual Real.* **2011**, *15*, 225–238. [CrossRef]
39. Bekele, M.K.; Pierdicca, R.; Frontoni, E.; Malinverni, E.S.; Gain, J. A survey of augmented, virtual, and mixed reality for cultural heritage. *ACM J. Comput. Cult. Herit.* **2018**, *11*, 7. [CrossRef]
40. Chen, J.; Wang, R. AI-augmented creative pipelines in small studios: A cost-benefit analysis. *J. Digit. Media Manag.* **2023**, *11*, 215–230.
41. He, Q.; Nguyen, L.T. AI-based multi-modal fusion for 3D face modeling from single-image recognition. In Proceedings of the 20th International Conference on Humanities and Social Sciences, Khon Kaen, Thailand, 7–8 January 2026; pp. 227–237. Available online: https://ichuso.kku.ac.th/paper/8 (accessed on 21 July 2026).

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.