Creation of 3D Image R&D

1) Quick problem framing & options

- **Goal:** produce a rotatable 3D asset (mesh + texture or neural render) from a user upload so it can be rotated/viewed interactively.
- **Three production paths:**

  1. **Single-image → single-view 3D (fast, lower fidelity):** depth- and priors-based (mesh + inferred texture) or learned single-view NeRFs.

  2. **Few-views (2–8 images / phone video):** multi-view stereo + neural rendering — much better fidelity.

  3. **Full multi-view (many photos / turntable):** classical SfM/MVS pipelines (COLMAP/OpenMVS) or NeRF with known poses — best fidelity.

2) Core pipeline components (typical)

1. **Preprocess & camera pose (if multi-view):** detect/extract EXIF, estimate poses (SfM) or use known capture rig. Tools: COLMAP for SfM/MVS. (GitHub)

2. **Depth / geometry estimation:** monocular depth (MiDaS/DPT) for single-image priors or MVS depth maps for multi-view. (GitHub)

3. **Surface reconstruction / mesh:** convert depth or SDF to mesh (Poisson, marching cubes, SDF extraction). For learned implicit surfaces use PIFu/PIFuHD (human-specific) or NeRF->mesh extraction via SDF/VolSDF. (arXiv)

4. **Texture / color:** project images onto mesh or "neural texture" from neural renderers (NeRF / DIB-R). Differentiable renderers such as DIB-R help training from 2D supervision. (NVIDIA)

5. **Retopology & UVs:** optional for real-time use — create lighter meshes and bake textures.

6. **Interactive viewer:** three.js / Babylon / custom WebGL; for NeRF-style results you may host a lightweight renderer (or bake views to a cubemap/video for very low latency).

3) Key families of AI models / techniques (with short notes & citations)

Neural Radiance Fields (NeRF) & variants

- Classic NeRF learns a volumetric radiance field from many posed images — photoreal but needs multiple views and time to optimize. Instant-NGP speeds training tremendously using a multiresolution hash encoding (very useful for interactive or fast turnaround). (nvlabs.github.io)

Single/few-image conditioned NeRFs

- **pixelNeRF** conditions a NeRF on one (or a few) input images so you can reconstruct with far fewer views — good compromise when you only have one photo. (arXiv)

Implicit surface methods for humans / clothed people

- **PIFu / PIFuHD** (pixel-aligned implicit functions) generate detailed meshes for clothed humans from a single image (very strong for people). (shunsukesaito.github.io)

Face-specific detailed reconstruction

- **DECA** (and related 3DMM-based methods) excel at detailed, animatable facial geometry from a single "in-the-wild" image — great if the use-case is heads/faces. (deca.is.tue.mpg.de)

Differentiable renderers / mesh-from-image

- **DIB-R** enables training mesh/texture predictors from only 2D supervision, helpful for single-image textured object prediction. (NVIDIA)

Classical multi-view pipelines

- **COLMAP** + MVS + OpenMVS: robust, proven pipelines for multi-photo reconstruction when camera poses can be estimated. Use these when you can ask users for multiple photos or a short turntable video. (GitHub)

Monocular depth priors

- **MiDaS / DPT**: state-of-the-art monocular depth estimators useful as priors for single-image mesh generation and to speed up hybrid pipelines.

4) Top recommended models (shortlist + why)

Top 5 picks (covering different use cases)

1. **Instant-NGP (NVLabs)** — *If you want fast, high-quality neural rendering / NeRF-based outputs and can ask for multiple views or short videos.* Instant training and rendering; excellent for prototyping photoreal rotatable views. (Requires CUDA

GPU.) (nvlabs.github.io)

2. **pixelNeRF** — *If you must support single-image inputs but want NeRF-quality renders.* Trains a model to condition a radiance field on a single image, enabling plausible novel views from one photo. Good for single-shot rotatable preview. (CVPR)

3. **PIFuHD** — *If your primary content is humans / clothed people from a single image.* Produces high-resolution mesh geometry (1k detail) tailored to humans — better than general single-image methods for people. (arXiv)

4. **DECA** — *If focus = faces/head reconstruction + animation.* Produces animatable, high-detail face geometry and albedo from one photo — ideal for avatars and expressive head models. (deca.is.tue.mpg.de)

5. **DIB-R / Differentiable rendering approaches** — *If you want to train models end-to-end from 2D images to textured meshes (particularly for objects).* Differentiable renderers enable supervision from images without 3D ground truth. (NVIDIA)

When to pick what

- Single person portrait → **DECA (face)** + **PIFuHD (full body)**
- General objects from single photo → **pixelNeRF** or DIB-R-based mesh+texture predictor.

5) R&D plan (phases and experiments)

Phase 0 — quick prototyping (2–4 weeks)

- Build minimal pipeline for *single-image* previews:
    - Run MiDaS to get depth map → generate point cloud → Poisson surface → basic texture projection (quick viewer). Use this to validate UX.
    - Parallel experiment: run a pre-trained pixelNeRF or small DIB-R model to produce novel views and compare visual quality/cost.

Phase 1 — focused model evaluation (4–8 weeks)

- Evaluate PIFuHD on human images (single view). Measure: geometry fidelity, artifacts, finger/clothing detail.

- Evaluate pixelNeRF vs. Instant-NGP with a few-view capture workflow — compare training time, visual quality, and memory/GPU needs. (nvlabs.github.io)

Phase 2 — production readiness (6–12 weeks)

- Add retopology / UV baking (for mobile/web realtime). Consider offline heavy compute to generate LODs and serve lightweight glTF/GLB.

- Build capture UX: single image vs guided multi-view capture. For multi-view, integrate on-device camera pose hints (ARKit/ARCore poses) to speed SfM/NeRF.

- Performance: use Instant-NGP or precompute baked textures/atlas for web delivery.

Phase 3 — robustness & dataset

- Collect a small dataset tailored to your target domain (humans, products, scenes). Evaluate metrics (Chamfer distance where ground truth exists, LPIPS for render similarity, perceptual user studies).

6) Datasets & evaluation suggestions

- **Datasets:** BlendedMVS / DTU for objects/multi-view; THuman / RenderPeople / 3D People datasets for human models; synthetic renders for ground truth. (Use these for quantitative tests and ablation studies.)

- **Metrics:** Chamfer distance (geometry), IoU/Normal consistency, LPIPS/PSNR for rendered images, and user perceptual evaluation for "rotatability" / realism.

7) Implementation & infra notes

- **GPU:** NeRF/Instant-NGP requires CUDA-capable GPUs; Instant-NGP uses tiny-cuda-nn and is extremely fast on modern NVIDIA hardware. (GitHub)

- **Latency strategy:** For real-time web viewing, precompute mesh + baked texture or produce a set of pre-rendered view images / light field. NeRFs are expensive to run in-browser unless heavily optimized or baked to impostors.

- **Open-source stacks:** COLMAP (SfM/MVS), MiDaS (depth), PIFu/PIFuHD (human mesh), pixelNeRF repos, Instant-NGP repo are available — good starting points.

8) Shortlist of repositories / papers to start cloning and testing

- Instant-NGP (NVlabs) — fast NeRF and primitives. (GitHub)

- pixelNeRF — conditioned NeRF for 1/few images. (GitHub)
- PIFuHD — single-image clothed human reconstruction (CVPR 2020). (shunsukesaito.github.io)
- DECA — detailed face reconstruction & animation. (deca.is.tue.mpg.de)
- COLMAP — SfM/MVS pipeline for multi-photo capture. (GitHub)
- DIB-R — differentiable renderer for single-image textured object learning. (NVIDIA)

9) Recommended next steps (actionable)

1. **Decide capture UX:** allow single photo only, or prompt for 3–8 extra photos / short turntable video? If you can get 4–8 images, quality jumps significantly.
2. **Prototype 2 flows in parallel**
   - *Single-image pipeline:* MiDaS → mesh → texture + pixelNeRF inference.
   - *Few-view pipeline:* instant-ngp with 6–12 phone frames (or COLMAP→MVS→mesh).
3. **Run a small user study** (10–20 images across subjects) to pick the best default flow.
4. **Optimize for delivery:** decide whether to deliver glTF/GLB (mesh + textures) or a lightweight neural viewer (for NeRF outputs).

10) Final recommendation — best models to try *first*

- **Start with Instant-NGP** for multi-view / short-video inputs (best quality/speed tradeoff). (nvlabs.github.io)
- **For single-photo human portraits**: PIFuHD (full body) + DECA (face detail). (CVPR)
- **For single-photo general objects**: evaluate pixelNeRF and a DIB-R style mesh+texture predictor. (CVPR)