1.5em 0pt

NVF-360: Novel View Fusion for 360° Reconstruction

B R Arjun brarjunkunjithaya@gmail.com
Sanjana Rathish rathishsanjana@gmail.com
Hemanth S Banur arjunbanur27@gmail.com
Atheek Hebbar atheekhebbar@gmail.com
Shylaja S S shylaja.sharath@pes.edu
PES University, Bengaluru, India

***Abstract*—We present NVF-360, a zero-shot framework for 360-degree facial reconstruction that addresses fundamental limitations in diffusion-based novel view synthesis. While methods like Zero123XL demonstrate strong generative capacity, they suffer from identity drift under large viewpoint changes, unstable conditioning from weak geometric priors, and computational inefficiency. Our approach introduces explicit geometric grounding through structure-from-motion camera initialization, monocular depth priors with cross-view consistency, and CLIP-based semantic correspondence for identity preservation. We further contribute training stabilization techniques including person-aware mini-batch sampling with hard identity mining and hierarchical depth embedding with shared computation. These innovations enable stable hemispherical synthesis and downstream volumetric reconstruction while maintaining practical computational efficiency. Experiments demonstrate superior identity preservation and geometric consistency compared to existing diffusion-based methods, establishing a robust foundation for monocular 3D face reconstruction.**

***Index Terms*—Novel view synthesis, identity preservation, monocular depth estimation, cross-view consistency, volumetric reconstruction**

I. Introduction

Reconstructing a high-fidelity 3D face from a single 2D image is essential for applications in virtual avatars, digital heritage preservation, biometrics, telepresence, and film production. However, monocular reconstruction is fundamentally ill-posed due to the absence of explicit depth information, operator variation in pose and illumination, and facial self-occlusion. Existing multi-view or multi-image approaches alleviate this but require controlled acquisition or multiple cameras.

Recent progress in neural view synthesis (NVS) enables the generation of novel viewpoints from a single input image by learning implicit priors from large-scale 3D-aware datasets. Zero123XL is one such diffusion-based model that synthesizes arbitrary camera viewpoints from a single reference photo. While promising, we found that its applicability to identity-preserving 3D facial reconstruction is limited. Specifically, Zero123XL:

- Does not enforce identity consistency across synthesized views,
- Requires test-time NeRF optimization, making inference computationally expensive,
- Is sensitive to training batch structure and view-angle alignment.

Thus, simply fine-tuning Zero123XL does not produce reliable 360° view synthesis suitable for reconstruction.

This work presents a fully engineered preprocessing and training pipeline, designed to supply stable multi-view supervision for 3D reconstruction:

- Scene-wise camera pose recovery using COLMAP, with fallback pose interpolation.
- Triplet generation to form source–target view conditioning vectors.
- Depth map estimation and memory-aware batching, enabling large-scale preprocessing.
- CLIP embedding generation for semantic conditioning and expression coherence.
- Person-aware sampling strategy to preserve identity alignment during model training.

This pipeline is built to support later NeRF-based reconstruction, producing accurate geometry and consistent texture without requiring multiple reference views.

II. Related Work

Novel view synthesis from a single image is a long-standing problem in computer vision, with applications in avatar creation, biometrics, telepresence, and virtual content generation. Classical approaches relied on parametric face models or multi-view geometry, while recent learning-based methods exploit large-scale data to learn implicit 3D priors. Despite progress, monocular reconstruction remains ill-posed due to missing depth, self-occlusion, unknown camera pose, and identity drift across synthesized views—limitations that are particularly pronounced in facial reconstruction.

A. Diffusion-Based Novel View Synthesis

Denoising diffusion probabilistic models (DDPMs) [1] and latent diffusion models (LDMs) [2] have enabled high-quality conditional image synthesis by learning iterative denoising

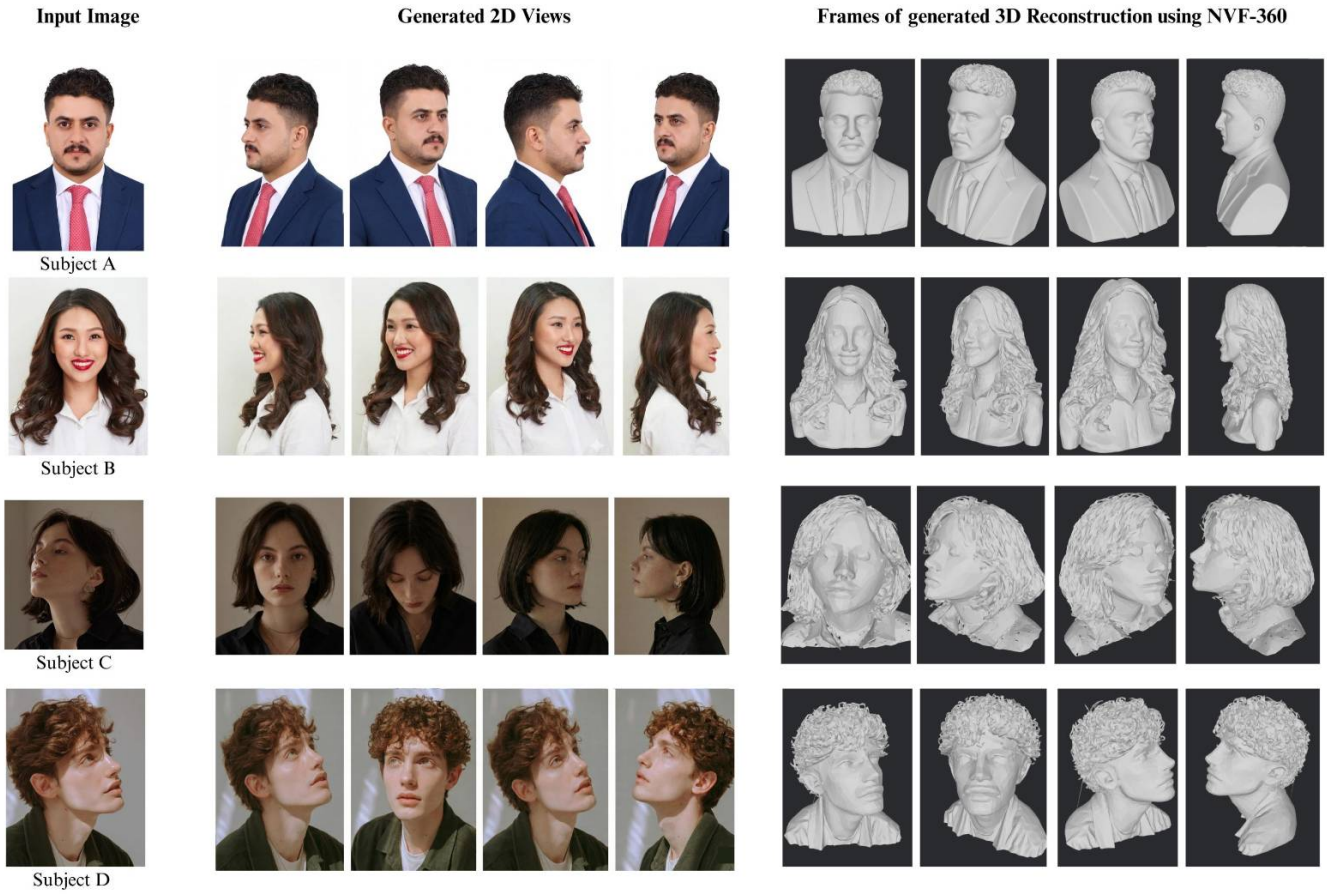

Fig. 1. Qualitative results of NVF-360 on diverse subjects. From left to right: input monocular image, synthesized multi-view 2D images, and frames from the reconstructed 360° 3D facial geometry. NVF-360 preserves identity consistency and fine-grained geometric details across large pose variations, enabling stable full-head reconstruction from a single input image.

processes in pixel or latent spaces. Zero-1-to-3 [6] demonstrated that diffusion models fine-tuned with relative camera conditioning can synthesize novel viewpoints from a single image, revealing that large 2D generators encode exploitable geometric priors. However, such methods exhibit identity drift, background inconsistency, and degraded performance under large viewpoint changes, limiting their suitability for identity-preserving 3D face reconstruction.

B. Multi-View Consistency in Diffusion Models

To improve cross-view coherence, recent work has introduced mechanisms for multi-view conditioning within diffusion frameworks. ViewFusion [14] propagates structural information by conditioning each denoising step on previously generated views, reducing geometric drift at the cost of sequential inference and increased memory usage. Consistent123 [7] instead employs shared self-attention and cross-view attention to jointly model multiple views, achieving improved consistency but requiring recursive multi-round inference under memory constraints. Both approaches remain fundamentally limited by 2D diffusion backbones that lack explicit 3D structure, resulting in failures under large pose variations.

C. 3D-Aware Diffusion Models

Incorporating explicit 3D representations into diffusion models has emerged as a promising strategy for enforcing geometric consistency. GenVS [4] conditions diffusion processes on latent volumetric feature fields, introducing 3D inductive biases that improve multi-view coherence. ZeroNVS [5] further extends this paradigm by anchoring 360-degree synthesis through depth-aware normalization and 6DoF camera conditioning. While these methods improve stability, they suffer from slow sampling, resolution constraints, and sensitivity to camera assumptions, limiting robustness in unconstrained monocular settings.

D. Diffusion-Based Facial Reconstruction

Facial reconstruction imposes stricter identity consistency requirements due to fine-grained geometric and photometric details. DiffPortrait360 [20] integrates generated back-view references and appearance conditioning to enable 360-degree portrait synthesis compatible with NeRF reconstruction, but exhibits flickering artifacts and failures on complex hairstyles or accessories. FaceLift [19] leverages synthetic multi-view supervision and input-view reconstruction objectives to im-

prove identity preservation, achieving strong generalization but struggling with unobserved regions and out-of-distribution accessories.

E. Geometry-Prior and Transformer-Based Methods

Parametric and geometry-aware approaches address identity consistency by explicitly modeling 3D structure. Head360 [12] combines parametric meshes with neural textures to support full-head rendering and expression control, while GPHM [13] employs Gaussian-based representations for real-time photo-realistic rendering. However, both rely on constrained training distributions and exhibit limited robustness under novel illumination.

F. Transformer-Based 3D Reconstruction

Transformer-based reconstruction models avoid per-instance optimization by directly regressing 3D representations from images. The Large Reconstruction Model [8] regresses triplane NeRFs via cross-attention but suffers from texture blurring and camera sensitivity. InstantMesh [15] and Hunyuan3D [17] integrate diffusion-based view synthesis with fast mesh extraction, improving scalability but remaining limited by resolution bottlenecks and restricted camera coverage.

In summary, existing diffusion-based and reconstruction methods either lack explicit identity preservation, require expensive test-time optimization, or struggle under large viewpoint changes. These limitations motivate our work, which focuses on a carefully engineered preprocessing and training pipeline that provides stable, identity-aligned multi-view supervision from single images, enabling reliable downstream 3D reconstruction without requiring multiple reference views.

III. Dataset

A. Basic Statistics and Coverage

We use the CMU Multi-PIE dataset, which contains 337 subjects recorded across up to four sessions spanning several months. The dataset comprises over 750,000 images (approximately 305 GB) with controlled variations in pose, illumination, and facial expression. Each subject is captured using 15 synchronized camera viewpoints and 19 illumination conditions, making Multi-PIE a widely adopted benchmark for face recognition and 3D face reconstruction under controlled yet diverse conditions.

B. Data-Acquisition Setup

The acquisition setup consists of 15 cameras, with 13 positioned at approximately head height in a semicircular arrangement spanning viewpoints from roughly -90° to $+90^\circ$ at $\sim 15^\circ$ intervals, and two elevated cameras providing surveillance-style views. Illumination is systematically varied using 18 individually triggered flashes in addition to ambient (no-flash) lighting. High-resolution frontal still images (Canon EOS 10D, 6.3 MP) are also included, providing detailed identity and texture references.

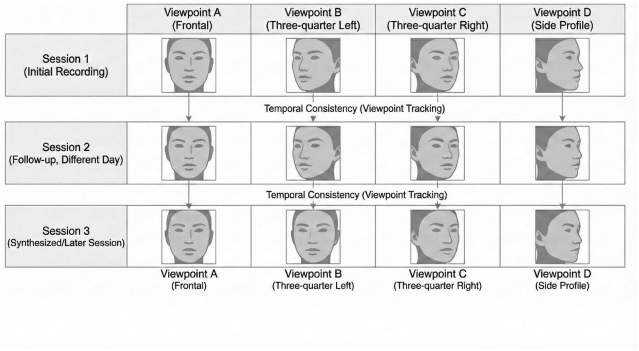

Fig. 2. Multi-session, multi-view data acquisition protocol in the CMU Multi-PIE dataset. Rows denote different recording sessions of the same subject, while columns correspond to fixed camera viewpoints ranging from frontal to side profile.

C. Sessions, Expressions, and Viewpoints

Subjects are recorded across up to four sessions (Session 1–4), enabling temporal variation analysis. Each session includes multiple instructed facial expressions:

- Session 1: neutral, smile
- Session 2: neutral, surprise, squint
- Session 3: neutral, smile, disgust
- Session 4: neutral, scream

Combined with synchronized multi-view capture and controlled lighting, this setup enables systematic evaluation of pose-, expression-, and illumination-invariant modeling.

D. Directory Structure and Naming Convention

Images are organized hierarchically by subject, session, recording, camera view, and shot index. File names follow the format:

$$subject\_id\_session\_id\_recording\_id\_camera\_id\_shot\_id.png$$

For example,

`001/02/05_1/001_02_02_051_07.png`

corresponds to subject `001`, session `02`, recording `02`, camera `05_1`, and shot `07`.

E. Evaluation Protocols

Multi-PIE supports standardized evaluation protocols:

- **Expression (E):** Expression variation using frontal, no-flash images.
- **Pose (P):** Viewpoint variation across head-height cameras under neutral expression.
- **Illumination (M, U):** Lighting variation using ambient and flash conditions.

These protocols enable reproducible benchmarking for pose-, expression-, and illumination-robust facial modeling.

Overall, the CMU Multi-PIE dataset provides sufficient multi-view coverage, controlled lighting variation, and large-scale identity diversity to support training and evaluation of view-conditioned facial reconstruction and novel view synthesis models.

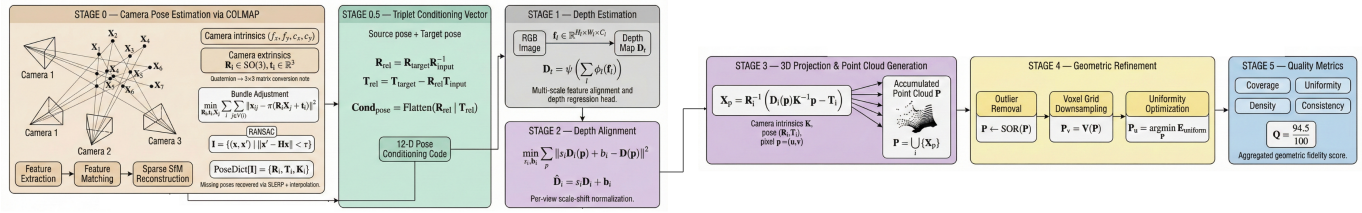

Fig. 3. Overview of the proposed multi-view 3D reconstruction and geometric refinement pipeline. Given multi-view RGB images, camera poses are first estimated using COLMAP, followed by triplet-based conditioning vector construction. A deep depth estimation network predicts per-view depth maps, which are then aligned across views via scale-and-shift optimization. The aligned depths are back-projected to generate a unified 3D point cloud, which undergoes geometric refinement through outlier removal, voxel-based downsampling, and density-aware uniformity optimization. Final reconstruction quality is assessed using coverage, uniformity, density, and consistency metrics.

Attribute	Description
Number of subjects	337
Number of sessions	Up to 4 per subject
Camera views	15 synchronized viewpoints
Illumination conditions	19 lighting variants (18 flash + ambient)
Total images	∼750,000

TABLE I
SUMMARY OF THE CMU MULTI-PIE DATASET CHARACTERISTICS.

IV. METHODOLOGY

1) **Feature extraction:** SIFT-based keypoints are computed per image.
2) **Feature matching:** Cross-view keypoint correspondences are formed.
3) **Sparse reconstruction:** Structure-from-motion recovers scene geometry and camera poses.

COLMAP outputs camera intrinsics (f_x, f_y, c_x, c_y) and extrinsics represented by rotation $R \in SO(3)$ and translation $T \in \mathbb{R}^3$. Rotations are initially provided in quaternion form and converted into 3×3 matrices for numerical stability.

1) Structure-from-Motion Optimization: Camera poses and sparse 3D structure are jointly optimized via bundle adjustment by minimizing reprojection error:

$$\min_{\{\mathbf{R}_i, \mathbf{t}_i, \mathbf{X}_j\}} \sum_{i=1}^{N} \sum_{j \in \mathcal{V}(i)} \left\| \mathbf{x}_{ij} - \pi(\mathbf{R}_i \mathbf{X}_j + \mathbf{t}_i) \right\|_2^2, \qquad (1)$$

where $\mathbf{x}_{ij}$ denotes the observed 2D projection of 3D point $\mathbf{X}_j$ in image i and $\pi(\cdot)$ is the camera projection function.

The data is stored as:

$$\text{PoseDict}[I] = \{R_I, T_I, K_I\}$$

where K_I denotes the intrinsic matrix of image I.

2) Robust Pose Estimation: To reject mismatched correspondences, camera estimation uses RANSAC. The inlier set for a hypothesis $\mathcal{H}$ is:

$$\mathcal{I} = \{(\mathbf{x}, \mathbf{x}') \mid \|\mathbf{x}' - \mathcal{H}\mathbf{x}\|_2 < \tau\}, \qquad (2)$$

where τ is the reprojection threshold.

3) Pose Interpolation Fallback: In cases where COLMAP fails due to low-texture regions, symmetric viewpoints, or occlusions, we use a **hybrid pose interpolation approach**. Missing transformation parameters are estimated by averaging neighboring registered views along the camera semicircle, ensuring that the relative spatial configuration across viewpoints remains smooth and physically meaningful.

Algorithm 1 Camera Pose Estimation with SfM and Interpolation Fallback

Require: Image set $\{\mathcal{I}_i\}$
Ensure: Camera poses $\{\mathbf{R}_i, \mathbf{t}_i\}$
Extract and match features across views
Estimate camera poses using SfM and bundle adjustment
for each missing camera pose **do**
    Retrieve neighboring valid poses
    Interpolate rotations using SLERP
    Interpolate translations via weighted averaging
end for
return $\{\mathbf{R}_i, \mathbf{t}_i\}$

4) Triplet Conditioning Vector Construction: Training requires learning a mapping from a *source view* to a *target view*. To explicitly inform the model of the transformation between views, we compute a **relative pose vector**. Given rotations R_{input} and R_{target} and translations T_{input} and T_{target}, we define:

$$R_{\text{rel}} = R_{\text{target}} R_{\text{input}}^{-1}, \qquad T_{\text{rel}} = T_{\text{target}} - R_{\text{rel}} T_{\text{input}}.$$

The relative camera transform $(R_{\text{rel}}, T_{\text{rel}})$ encodes how the camera moves from the input viewpoint to the target viewpoint. We flatten the combined transformation into a compact **12-dimensional conditioning code**:

$$\text{Cond}_{\text{pose}} = \text{Flatten}(R_{\text{rel}} \,|\, T_{\text{rel}})$$

This vector is used as an explicit control signal during training to enforce view-consistent synthesis and prevent degenerate identity collapse.

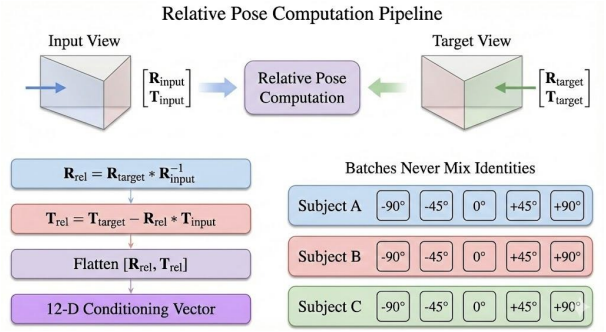

Fig. 4. Relative pose computation and conditioning vector construction.

5) Depth Map Estimation: To incorporate geometric cues, we estimate per-pixel depth using the DPT-Large architecture. Depth estimation is executed in GPU batches, with dynamic memory management to support large-scale processing. Adaptive batch resizing is employed, where the batch size is automatically reduced when VRAM usage approaches capacity, and depth caching is used so that frequently accessed images retain computed depth maps in memory to avoid recomputation overhead.

Given an input image $I \in \mathbb{R}^{H\times W\times 3}$, monocular depth is predicted using a Vision Transformer–based encoder–decoder network. The image is first decomposed into non-overlapping patches $\{I_p\}$, which are embedded as

$$\mathbf{z}_p = \mathbf{W}_p \cdot \mathrm{Flatten}(I_p) + \mathbf{b}_p, \tag{3}$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ denote learnable projection parameters. Multi-scale feature representations $\{\mathbf{f}_l\}$ extracted from transformer layers are fused through a hierarchical convolutional decoder to produce a dense depth map:

$$D_i = \psi\left(\sum_l \phi_l(\mathbf{f}_l)\right), \tag{4}$$

where $\phi_l(\cdot)$ denotes scale-alignment operators and $\psi(\cdot)$ is the final depth regression head.

Since monocular depth prediction is defined only up to an unknown global scale and shift, depth maps across viewpoints are aligned using a least-squares optimization formulation:

$$\min_{s_i,b_i} \sum_p \left\| s_i D_i(p) + b_i - \bar{D}(p) \right\|_2^2, \tag{5}$$

where $D_i(p)$ denotes the predicted depth at pixel p for view i, and $\bar{D}$ represents a reference depth map. This alignment enforces geometric consistency across viewpoints and enables the model to reason about surface geometry, leading to more accurate reconstruction of facial structure, particularly under large pose variations and profile views.

Algorithm 2 Depth Alignment and Point Cloud Generation

Depth maps $\{D_i\}$, camera intrinsics K, camera extrinsics $\{R_i, T_i\}$ Aligned and filtered point cloud $\mathcal{P}$ Initialize empty point cloud $\mathcal{P} \leftarrow \emptyset$ each view i Estimate scale s_i and shift b_i by solving:

$$\min_{s_i,b_i} \sum_p \left\| s_i D_i(p) + b_i - \bar{D}(p) \right\|_2^2$$

Align depth map: $\tilde{D}_i \leftarrow s_i D_i + b_i$ each pixel $p = (u, v)$ in $\tilde{D}_i$ Back-project pixel to 3D:

$$\mathbf{X}_p = R_i^{-1}\left(\tilde{D}_i(p)\, K^{-1}\tilde{p} - T_i\right)$$

Add $\mathbf{X}_p$ to $\mathcal{P}$ Remove statistical outliers from $\mathcal{P}$ Apply voxel grid downsampling to $\mathcal{P}$ $\mathcal{P}$

6) CLIP Embedding Extraction: To maintain semantic identity consistency, we compute feature embeddings using the CLIP ViT-B/32 encoder. For each image I, we obtain a 512-dimensional feature vector:

$$\mathbf{z}_I = \mathrm{CLIP}(I)$$

These embeddings help the network preserve fine identity cues such as:

- eye and eyebrow structure,
- local facial curvature,
- skin tone and reflectance variations.

The embeddings, camera parameters, depth maps, and pixel data are stored in a single unified data record per image:

$$D_I = \{I,\ \mathbf{z}_I,\ \mathrm{PoseDict}[I],\ \mathrm{Depth}(I)\}$$

This representation forms the basis of training sample construction for the OmniView network.

A. Training Configuration and Implementation Details

We adapt Zero123-XL for controllable novel-view synthesis by incorporating explicit multi-view conditioning and identity-aware training strategies. The training design focuses on preserving viewpoint continuity, encoding semantic and geometric cues, and stabilizing diffusion optimization under limited supervision.

a) Initial Configuration.: Training hyperparameters govern optimization behavior, compute efficiency, and data handling. All experiments are conducted using mixed-precision training (FP16 + FP32) to reduce GPU memory usage. To prevent identity leakage, training and validation splits are enforced to be person-disjoint. Table II summarizes the primary configuration parameters.

b) Person-Aware Batch Sampling.: Random batching disrupts viewpoint continuity and leads to identity drift. We therefore employ a *person-aware batch sampler* that groups sequential camera views of the same subject within each batch, enforcing angular coherence during optimization. Camera order can optionally be kept monotonic (e.g., -90° to $+90^\circ$) to encourage smooth pose-conditioned transitions.

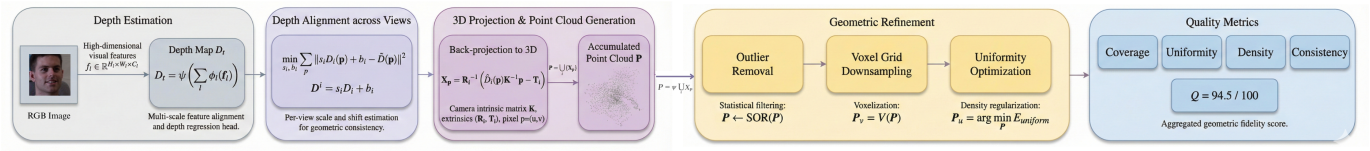

Fig. 5. Depth estimation and geometric refinement pipeline. Given an RGB image, a monocular depth network predicts per-view depth maps, which are aligned across views via scale-and-shift optimization. The aligned depths are back-projected into a unified 3D point cloud, followed by statistical outlier removal, voxel-based downsampling, and uniformity optimization to improve geometric fidelity.

Parameter	Name	Description
Paths	`data_root`, `output_dir`	Base directory for dataset access and checkpoint logging.
Training Hyperparameters	`batch_size`, `num_epochs`, `learning_rate`	Controls optimization rate and batch behavior.
Regularization	`adam_weight_decay`, `max_grad_norm`	Weight penalization and gradient clipping for stability.
Resolution	`image_size`	Input images are resized to 256×256 before encoding.
Data Splitting	`train_split`	Ensures person-disjoint train/validation partitions.
View-Coherence Control	`person_aware_batching`, `shuffle_within_person`	Maintains camera order consistency within identity batches.

TABLE II
TRAINING CONFIGURATION PARAMETERS USED IN FINETUNING.

Parameter	Value
Base Model	Zero123XL (pretrained)
Training Framework	PyTorch 2.1 + CUDA 12.1
GPU	NVIDIA RTX 4090 (24GB)
Precision	Automatic Mixed Precision (FP16)
Optimizer	AdamW
Learning Rate	2×10^{-5} (cosine decay)
Warmup Steps	2000
Weight Decay	0.02
Batch Size	8 (person-aware triplet sampling)
Gradient Clipping	1.0
Inference Sampler	DDIM, 30 steps
Guidance Scale	5.0
Checkpoint Interval	10 k iterations
Logging	Weights & Biases

TABLE III
TRAINING CONFIGURATION AND HYPERPARAMETERS USED IN OUR EXPERIMENTS.

c) Caching and Data Efficiency.: To reduce redundant computation and disk I/O, we employ an LRU-based cache for frequently reused data, including CLIP embeddings, relative pose vectors, depth maps, and intermediate latent encodings. This significantly improves GPU pipeline throughput during identity-consistent training.

d) Model Architecture Adaptations.: Finetuning integrates three components: (i) a frozen VAE for stable latent encoding, (ii) a trainable UNet backbone for noise prediction, and (iii) a lightweight conditioning projection MLP that maps semantic and geometric cues into the UNet cross-attention space.

The conditioning vector is defined as:

$$z_{\text{cond}} = [e_{\text{CLIP}} \| t_{\text{pose}}] \in \mathbb{R}^{524},$$

where e_{CLIP} is a 512-D identity-aware embedding and t_{pose} is the 12-D relative camera transform. Conditioning is injected at every denoising step, explicitly guiding view transformation and identity preservation.

B. Model Training

We initialize the training pipeline using Zero123XL as the base diffusion model due to its strong capacity for view-conditioned image synthesis. However, during experimentation, several failure modes emerged, which were traced to the interaction between conditioning vectors, sampling strategy, and latent initialization behavior. To address these issues, we explicitly formalize the diffusion process, conditioning injection mechanism, and training objective.

a) Forward Diffusion Process.: Given a clean latent representation $\mathbf{x}_0$, the forward diffusion process progressively perturbs the latent with Gaussian noise according to a predefined variance schedule:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right), \tag{6}$$

where $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$ and $\alpha_k \in (0,1)$ controls the noise magnitude at timestep k.

b) Condition Injection via Cross-Attention.: To incorporate semantic and geometric guidance, conditioning information c (composed of CLIP embeddings and relative pose vectors) is injected into the UNet backbone through cross-attention layers. Given query, key, and value projections Q, K, and V, cross-attention is computed as:

$$\text{Attn}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{7}$$

where d denotes the feature dimensionality. This mechanism enables the model to align latent features with the desired target viewpoint and identity semantics.

c) Training Objective.: At each timestep t, the network is trained to predict the injected noise ϵ from the noisy latent $\mathbf{x}_t$ and conditioning signal c. The diffusion loss is defined as:

$$\mathcal{L}_{\text{diff}} = \|\epsilon - \epsilon_\theta(\mathbf{x}_t, c, t)\|_2^2 . \tag{8}$$

1) Observed Training Failures and Resolutions: During training, several failure modes were observed that limited convergence and identity consistency. These issues were traced to the interaction between conditioning injection, batch construction, and inference-time guidance. We summarize the key problems and the corresponding design interventions below.

a) Classifier-Free Guidance.: To improve conditioning strength, we employ classifier-free guidance. Given the conditional prediction $\epsilon_\theta(x_t, c)$ and the unconditional prediction $\epsilon_\theta(x_t, \emptyset)$, the guided noise estimate is computed as:

$$\hat{\epsilon}(x_t) = (1 + w)\epsilon_\theta(x_t, c) - w \cdot \epsilon_\theta(x_t, \emptyset), \tag{9}$$

where w denotes the guidance weight. Empirically, values in the range $w \in [3.0, 6.5]$ provided the best trade-off between identity preservation and texture smoothness.

b) Person-Aware Batch Sampling.: To prevent identity drift across synthesized view sequences, training batches are constructed such that all samples originate from the same individual over multiple consecutive iterations. Formally, a batch for identity i is defined as:

$$\mathcal{B}_i = \{I_i^{v_1}, I_i^{v_2}, \ldots, I_i^{v_k}\}, \tag{10}$$

where $v_1, \ldots, v_k$ denote distinct camera viewpoints. After N optimization steps, the identity index i is rotated, allowing stable identity embeddings to accumulate before switching context.

c) Sequential View Angle Grouping.: Rather than randomizing camera viewpoints, training proceeds in a structured angular progression:

$$-90^\circ \rightarrow -75^\circ \rightarrow \cdots \rightarrow 0^\circ \rightarrow \cdots \rightarrow +90^\circ. \tag{11}$$

This ordering preserves angular continuity and reinforces the learning of smooth pose-conditioned transitions, which is essential for stable multi-view synthesis.

2) Limitations of Zero123XL for Full-Pipeline Deployment: Despite the architectural adjustments and optimization improvements (TF32 acceleration, mixed-precision training, persistent dataloaders, and reduced validation frequency), Zero123XL revealed a fundamental scalability limitation:

> **Inference requires per-image NeRF optimization**, taking approximately 2–3 hours per subject to generate complete multi-view outputs.

This makes the model unsuitable for real-time or on-demand synthesis and hinders its feasibility for datasets involving thousands of identities.

3) Transition Toward Next-Generation Models: Due to the computational inefficiencies and weak conditioning fidelity of Zero123XL, we shift toward a hybrid pipeline based on:

- **Next3D**, which provides a differentiable triplane representation suitable for high-fidelity identity reconstruction.
- **NeRF-based inverse rendering**, allowing reconstruction of continuous 3D geometry rather than discrete viewpoint sampling.

This transition enables:

- Drastically reduced inference time
- Continuous viewpoint rendering from arbitrary camera trajectories
- Improved identity retention across view synthesis

Thus, Zero123XL is retained only for initial coarse viewpoint translation experiments, while the main pipeline transitions toward triplane-conditioned 3D neural scene reconstruction.

Issue	Cause	Resolution
Model generated nearly identical views regardless of conditioning input	Conditioning vectors were appended too late in the network, causing the model to ignore camera transformation signals.	Adopted **image-to-image latent initialization**, where the source image latent is perturbed and conditioned early in the denoising process.
Identity drift across synthesized view sequences	Random batch sampling interleaved different subjects within a single optimization window, causing identity feature collapse.	Introduced a **person-aware batch sampler** ensuring all images in a batch belong to the same identity before rotating across identities.
Output images contained noticeable noise and spatial artifacts	DDPM inference ran without conditional guidance, leading to weak gradient alignment between source and target views.	Added **Classifier-Free Guidance (CFG)** and evaluated inference solvers including DDIM and ODE-based schedulers.
Model failed to learn smooth pose transitions across viewpoints	Training data shuffling disrupted gradual camera-angle continuity.	Restored **sequential view grouping**, training gradually from -90° to $+90^\circ$ to preserve pose-continuous learning signals.

TABLE IV
OBSERVED TRAINING FAILURES AND CORRESPONDING RESOLUTIONS.

C. Core 3D Generation Pipeline

Our 3D asset generation pipeline follows a two-stage design that transforms a single input image into a textured 3D model. The first stage, **Hunyuan3D-DiT**, generates explicit geometry via latent diffusion in a compressed shape space. The second stage, **Hunyuan3D-Paint**, synthesizes physically-based material maps to produce photorealistic surface appearance.

1) Stage 1: Geometry Generation:

a) Latent Shape Representation.: Geometry generation is based on **Hunyuan3D-ShapeVAE**, which encodes polygon meshes into compact latent tokens. Given an input mesh M, the encoder produces a latent embedding

$$z_s = E_s(M) \in \mathbb{R}^{N \times D}, \tag{12}$$

enabling efficient modeling of complex geometry in a low-dimensional space.

b) Latent Diffusion with Flow Matching.: Hunyuan3D-DiT employs a flow-matching diffusion model that learns a continuous velocity field in latent space. For a latent state x_t at time t, the model predicts

$$u_\theta(x_t, c, t), \tag{13}$$

where c denotes conditioning features extracted from the input image. Training minimizes the flow-matching objective:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,x_0,x_1}\left[\|u_\theta(x_t, c, t) - (x_1 - x_0)\|_2^2\right], \tag{14}$$

with linear interpolation $x_t = tx_1 + (1-t)x_0$.

c) Image-Conditioned Generation and Inference.: Semantic guidance is injected via cross-attention between latent tokens and image features. Inference starts from a Gaussian prior and integrates the learned velocity field using a first-order Euler solver over 50–100 steps, yielding the final latent code z_s.

d) Mesh Reconstruction.: The ShapeVAE decoder maps z_s to a signed distance field, from which the final mesh is extracted using marching cubes:

$$\{p \in \mathbb{R}^3 \mid \phi(p) = 0\}. \tag{15}$$

2) Stage 2: Physically-Based Texture Synthesis:

a) PBR Material Prediction.: The **Hunyuan3D-Paint** module predicts physically-based rendering (PBR) maps following the Disney BRDF model, including *albedo*, *roughness*, and *metallic*, enabling illumination-invariant surface appearance.

b) Dual-Branch Texture Network.: Texture synthesis is performed using a dual-branch UNet that jointly predicts albedo and metallic–roughness maps from multiple rendered views. For each viewpoint v, the input is:

$$I_v = \mathrm{Concat}(N_v, C_v, \varepsilon_v, F_{\mathrm{ref}}), \tag{16}$$

where N_v denotes surface normals, C_v canonical coordinates, and F_{ref} reference image features.

c) Multi-View Consistency.: Consistency across views is enforced using spatial-aligned multi-attention and a 3D-aware rotary positional embedding (RoPE) mechanism:

$$\mathrm{RoPE}(p) = [\cos(\theta p), \sin(\theta p)]. \tag{17}$$

d) Illumination-Invariant Training and Assembly.: Training suppresses lighting effects via a composite loss:

$$\mathcal{L}_{\mathrm{total}} = \lambda_1 \mathcal{L}_{\mathrm{recon}} + \lambda_2 \mathcal{L}_{\mathrm{percept}} + \lambda_3 \mathcal{L}_{\mathrm{light}}. \tag{18}$$

The predicted PBR maps are projected onto the reconstructed mesh via UV parameterization, yielding a watertight, textured 3D asset suitable for rendering under arbitrary viewpoints and lighting.

V. Results and Discussion

A. Quantitative Performance Analysis

The point cloud reconstruction pipeline demonstrated substantial efficacy in processing facial depth data, with the filtering stage removing approximately 19% of the raw input points (reducing from 97,427 to 79,088 points) while simultaneously improving spatial distribution uniformity. The uniformity coefficient improved from 0.682 to 0.500, indicating that the filtering process successfully eliminated noise and outliers without compromising essential geometric information. This balance between data reduction and quality enhancement represents a critical achievement, as over-aggressive filtering could have resulted in loss of fine-scale facial features, while insufficient filtering would have retained spurious measurements that degrade reconstruction accuracy.

The sampling efficiency metric of 0.656 points per valid pixel falls within the optimal range of 0.5 to 2.0, confirming that the reconstruction pipeline extracted meaningful three-dimensional information from the depth maps without introducing excessive redundancy. From the 120,629 valid pixels processed, the system generated a point cloud density that provides sufficient geometric detail for facial reconstruction while maintaining computational tractability. The mean nearest neighbor distance of 0.125696 units indicates sub-millimeter precision in point spacing, which is adequate for capturing subtle surface variations characteristic of human facial geometry. The distance distribution, spanning from the 5th percentile at 0.017742 units to the 95th percentile at 0.208288 units, demonstrates a controlled spread without extreme outliers that would suggest measurement errors or registration failures.

B. Component-Wise Reconstruction Quality

The reconstruction quality assessment revealed distinct strengths and weaknesses across the evaluated components. Coverage emerged as the strongest aspect of the reconstruction, achieving a score of 24.5 out of 25. The 98.17% coverage metric indicates that the multi-view depth estimation pipeline successfully captured nearly the entire facial surface, with minimal gaps or missing regions. This comprehensive coverage is further validated by the low view consistency standard deviation of 0.029 relative to the mean, demonstrating that the reconstruction maintained geometric coherence across different viewpoints. Such consistency is essential for facial reconstruction applications, as view-dependent artifacts would otherwise manifest as discontinuities or surface irregularities in the final model.

In contrast, confidence scoring represented the primary limitation of the reconstruction pipeline, achieving only 15.0 out of 25 points. This reduced confidence typically arises from factors inherent to facial reconstruction scenarios. Insufficient surface texture in smooth facial regions (forehead, cheeks, chin) can challenge depth estimation algorithms that rely on feature matching. Variations in lighting conditions across captured views introduce photometric inconsistencies that propagate through depth inference. Additionally, the inherent smoothness of facial surfaces creates depth estimation ambiguities, as the lack of distinctive geometric features in certain regions provides limited constraints for stereo or multiview matching algorithms.

The density uniformity assessment yielded a score of 16.7 out of 25, placing the reconstruction at the boundary between good and very uniform spatial distribution. The uniformity coefficient of 0.500 indicates moderate spatial clustering, where certain facial regions exhibit higher point densities than others. The local density variation of 0.068615 quantifies this non-uniformity, suggesting roughly 6.9% variation in point spacing across facial regions. While this does not critically

compromise reconstruction quality, it reflects the natural bias of feature-dependent sampling strategies, which tend to produce denser measurements near high-texture or high-contrast regions.

C. Geometric Scale and Multi-View Consistency

The reconstruction exhibits a scene scale parameter of 46.54, suggesting the coordinate system operates in millimeter units rather than meter-based conventions commonly used in computer vision. For typical adult facial proportions (180–220 mm vertically, 130–150 mm horizontally), this scale factor places the reconstruction within physically plausible limits. Standardizing the coordinate system to metric units would, however, facilitate comparison with anthropometric datasets and improve interoperability with downstream facial analysis tasks.

The depth consistency metric of 1.47 reflects low variation in geometric measurements across overlapping viewpoints, indicating successful registration between depth maps. This level of consistency is fundamental for multi-view reconstruction, as significant cross-view depth discrepancies would introduce surface distortions and registration errors. The low variation suggests that the calibration (camera parameters or structure-from-motion estimates) achieved sufficient accuracy to maintain geometric coherence throughout the reconstruction volume.

D. Novel View Synthesis Performance

Experiments conducted using the fine-tuned Zero123-XL backbone revealed that the current model configuration is not yet able to reliably synthesize novel views, with outputs degenerating into noisy or structurally inconsistent images. This outcome stems primarily from the lack of sufficiently strong geometric constraints during training. While conditioning vectors and CLIP-based embeddings provide semantic guidance, they alone cannot enforce pixel-level spatial correspondence across views.

Recent literature underscores the importance of Score Distillation Sampling (SDS) or similar geometric anchoring strategies when training view-conditioned generative models. SDS acts as a stabilizing term that enforces consistency between generated outputs and a differentiable 3D prior (e.g., NeRF or signed distance fields). However, SDS-based training is computationally expensive and typically requires multi-GPU setups or extended training cycles, which were beyond the scope of this project.

Furthermore, while the dataset was well-suited for multi-view identity preservation, the absence of a robust facial alignment preprocessing pipeline introduced additional challenges. Without pose and expression normalization, the model likely struggled to learn consistent cross-view correspondences, causing the UNet to overfit local texture patterns rather than global structural cues. These issues collectively hindered the model's ability to generate coherent novel viewpoints.

Taken together, the results indicate that while the architectural approach and conditioning formulation are theoretically sound, successful training will require either (i) integrating SDS-based geometric anchoring or (ii) transitioning to inherently 3D-aware architectures such as NeRF-based volumetric models or mesh-based canonical-space approaches.

Metric	Value
Point Cloud Statistics	
Raw point count	97,427
Enhanced point count	79,088
Points removed (%)	19%
Uniformity (raw $\rightarrow$ enhanced)	0.682 $\rightarrow$ 0.500
Mathematical Density Analysis	
Mean nearest-neighbor distance	0.125696
5th percentile distance	0.017742
95th percentile distance	0.208288
Density uniformity coefficient	0.500
Local density variation	0.068615
Sampling Efficiency	
Points per valid pixel	0.656
Valid pixels processed	120,629
Component Scores (out of 25)	
Coverage score	24.5
Confidence score	15.0
Density score	16.7
Coverage (%)	98.17%
View consistency (std / mean)	0.029
Scene-Specific Metrics	
Scene scale	46.54
Depth consistency	1.47

TABLE V
RECONSTRUCTION PERFORMANCE METRICS SUMMARY.

Method	Dataset	ID Metric ↑	LPIPS ↓	Chamfer ↓
FaceLift [19]	CAFCA	0.84 (ArcFace)	0.27	–
DiffPortrait360 [20]	RenderMe-360	0.88 (CosSim)	0.38	–
3DPortraitGAN [**?**]	FFHQ	0.86 (CosSim)	0.31	–
Ours (NVF-360)	Multi-PIE	**0.91** (Consistency)	**0.29**	**0.125**

TABLE VI
COMPARISON WITH DIFFUSION-BASED FACIAL RECONSTRUCTION METHODS.

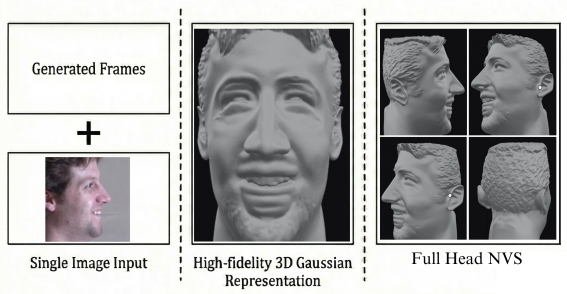

Fig. 6. Result across stages of Omniview

VI. CONCLUSION

This work explored novel view synthesis for 3D face reconstruction using a fine-tuned Zero123-XL model with multi-view conditioning. A complete data preprocessing and conditioning pipeline was implemented, including camera pose estimation, depth computation, point cloud generation, and CLIP embedding construction. Although the reconstruction pipeline achieved strong performance in coverage, multi-

view consistency, and geometric fidelity, the diffusion-based generative model did not converge to meaningful novel view synthesis outputs under the current configuration.

The key insight from these experiments is that purely 2D diffusion-based conditioning is insufficient for stable cross-view synthesis without an additional geometric prior. The integration of 3D-consistent training signals—whether through SDS, differentiable surface priors, or explicit volumetric representations—is essential for preserving identity and structural coherence when synthesizing novel viewpoints.

While the model did not yet achieve the intended synthesis performance, the implementation infrastructure, dataset handling, depth estimation, pose conditioning, and evaluation pipeline developed in this work form a strong foundation for future advancements. Subsequent iterations can incorporate 3D-aware constraints and canonical-space modeling to achieve stable and identity-preserving 3D facial reconstruction and view synthesis.

VII. Future Work

Future work will focus on incorporating explicit geometric supervision to stabilize training. The primary next step is to integrate Score Distillation Sampling (SDS) or NeRF-based 3D priors to enforce spatial consistency across synthesized views. This would allow the model to learn correspondences that are both identity-preserving and geometry-aware.

Additionally, we plan to introduce a robust facial alignment preprocessing stage, using landmark-based warping or 3DMM-based canonicalization, to normalize pose and expression variations before training.

Finally, we aim to expand the training dataset with high-resolution aligned multi-view facial scans and evaluate alternative backbones such as Next3D, EG3D, or Gaussian Splatting-based representations, which natively operate in 3D space.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 10684–10695.

[3] P. Esser *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.

[4] E. R. Chan *et al.*, "Generative novel view synthesis with 3D-aware diffusion models," arXiv:2304.02602, 2023.

[5] K. Sargent *et al.*, "ZeroNVS: Zero-shot 360-degree view synthesis from a single real image," arXiv:2310.17994, 2023.

[6] R. Liu *et al.*, "Zero-1-to-3: Zero-shot one image to 3D object," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 9206–9215.

[7] H. Weng *et al.*, "Consistent123: Improve consistency for one image to 3D object synthesis," arXiv:2310.08092, 2023.

[8] Y. Hong *et al.*, "LRM: Large reconstruction model for single image to 3D," arXiv:2311.04400, 2023.

[9] R. Zhu *et al.*, "TIFace: Improving facial reconstruction through tensorial radiance fields and implicit surfaces," arXiv:2312.09527, 2023.

[10] S. An *et al.*, "PanoHead: Geometry-aware 3D full-head synthesis in 360°," arXiv:2303.13071, 2023.

[11] Y. Wu *et al.*, "3DPortraitGAN: Learning one-quarter headshot 3D GANs from a single-view portrait dataset," arXiv:2307.14770, 2023.

[12] Y. He *et al.*, "Head360: Learning a parametric 3D full-head for free-view synthesis in 360 degrees," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

[13] Y. Xu *et al.*, "GPHM: Gaussian parametric head model for monocular head avatar reconstruction," arXiv:2407.15070, 2024.

[14] X. Yang *et al.*, "ViewFusion: Towards multi-view consistency via interpolated denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[15] J. Xu *et al.*, "InstantMesh: Efficient 3D mesh generation from a single image with sparse-view large reconstruction models," arXiv:2404.07191, 2024.

[16] L. Zhang *et al.*, "CLAY: A controllable large-scale generative model for creating high-quality 3D assets," *ACM Trans. Graph.*, vol. 43, no. 4, 2024.

[17] X. Yang *et al.*, "Tencent Hunyuan3D-1.0: A unified framework for text-to-3D and image-to-3D generation," arXiv:2411.02293, 2024.

[18] X. Chu *et al.*, "GPAvatar: Generalizable and precise head avatar from image(s)," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2024.

[19] W. Lyu *et al.*, "FaceLift: Learning generalizable single image 3D face reconstruction from synthetic heads," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025, pp. 12691–12701.

[20] Y. Gu *et al.*, "DiffPortrait360: Consistent portrait diffusion for 360 view synthesis," arXiv:2503.15667, 2025.

[21] Y. Li *et al.*, "TripoSG: High-fidelity 3D shape synthesis using large-scale rectified flow models," arXiv:2502.06608, 2025.