NVF-360: NOVEL VIEW FUSION FOR 360 RECONSTRUCTION

B R Arjun, Sanjana Rathish, Hemanth S Banur, Atheek Hebbar, Shylaja S S

PES University, Bengaluru
India

ABSTRACT

We present NVF-360, is a zero-shot, 360-degree facial reconstruction model which overcomes some of the core drawbacks of diffusion-based synthesis of novel views. Already existing approaches, such as Zero123XL, have a good generative ability, but they are affected by identity drift with large viewpoint variations, unstable conditioning with weak geometry priors, and have poor computational efficiency. Our method proposes clear geometrical grounding with structure-from-motion camera first-person, monocular depth priors with inter-view consistency and CLIP-based semantic correspondence to achieve identity preservation. We also add training stabilization methods such as person-aware mini-batch sampling and hard identity mining and hierarchical depth embedding with shared computation. These novel innovations allow the synthesis of stable hemispherical and downstream volumetric reconstructions, whilst retaining experimentally viable computational performance. Experiments show that in comparison to current diffusion-based methods, it achieves better identity-preservation and geometric-consistency, thereby creating a solid foundation for further monocular 3D face reconstruction.

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

1. INTRODUCTION

Reconstructing high-fidelity 3D faces from single 2D images is challenging due to missing depth, pose ambiguity, and self-occlusion. While multi-view methods alleviate these issues, they require controlled capture. Neural view synthesis models such as Zero123XL generate novel views from one image but suffer from identity drift, batch sensitivity, and costly test-time NeRF optimization. We propose a stable multi-view supervision pipeline that integrates COLMAP-based pose recovery, depth estimation, CLIP semantic conditioning, and person-aware sampling, enabling identity-consistent NeRF-based reconstruction from a single reference image.

2. RELATED WORK

Novel view synthesis from single images addresses 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 for implicit 3D priors. Monocular reconstruction remains ill-posed due to missing depth, self-occlusion, unknown camera pose, and identity drift, particularly pronounced in facial reconstruction.

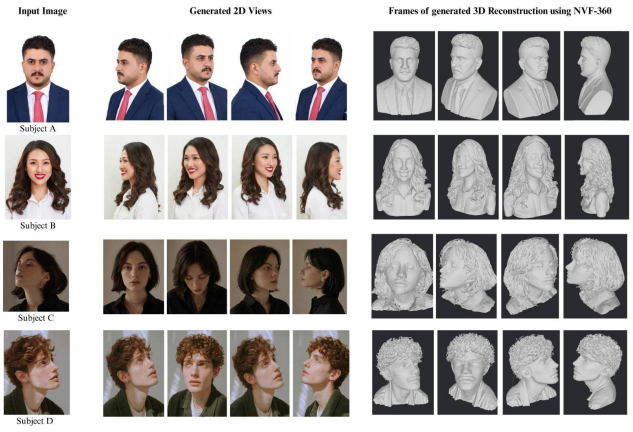

Fig. 1. Qualitative results of NVF-360 on diverse subjects. From left to right: input image, synthesized multi-view images, and views of the reconstructed 360° 3D face. NVF-360 preserves identity and fine-grained geometry across large pose variations from a single image.

Denoising diffusion probabilistic models and latent diffusion models [11] enable high-quality conditional synthesis through iterative denoising. Zero-1-to-3 [8] demonstrated diffusion models fine-tuned with relative camera conditioning can synthesize novel viewpoints, revealing exploitable geometric priors in 2D generators. However, these methods exhibit identity drift, background inconsistency, and degraded performance under large viewpoint changes.

Recent work introduces multi-view conditioning mechanisms. ViewFusion [18] propagates structural information by conditioning each denoising step on previously generated views, while Consistent123 [13] employs shared self-attention and cross-view attention to jointly model multiple views. Both remain limited by 2D diffusion backbones lacking explicit 3D structure, causing failures under large pose variations.

Incorporating explicit 3D representations enforces geo-

Attribute	Description
Number of subjects	337
Number of sessions	Up to 4 per subject
Camera views	15 synchronized viewpoints
Illumination conditions	19 lighting variants
Total images	~750,000

Table 1. CMU Multi-PIE dataset characteristics.

metric consistency. GenVS [2] conditions diffusion on latent volumetric feature fields, while ZeroNVS [12] anchors 360-degree synthesis through depth-aware normalization and 6DoF camera conditioning. These methods suffer from slow sampling, resolution constraints, and sensitivity to camera assumptions.

More rigid consistency of identity is needed in facial reconstruction. DiffPortrait360 [5] uses generated references of the back view and appearance conditioning to synthesize 360 portraits but contains flicker effects on complicated hairstyles. FaceLift [10] is an identity preservation system based on synthetic multi-view supervision that is faced with difficulties in unobserved regions. Parametric systems such as Head360 [6] combine meshes with neural textures for full-head rendering, and GPHM [15] uses Gaussian representations with constrained training distributions, use limited robustness under novel illumination are claimed to be applicable to full-head representations. Models that use cross-attention to regress triplane NeRFs, such as LRM [7], are at risk of losing texture and are not as scalable as diffusion-based view synthesis, while other models, such as InstantMesh [16] and Hunyuan3D, have adopted diffusion-based view synthesis with fast mesh extraction, improving scalability but limited by resolution bottleneck.

3. DATASET

We use the CMU Multi-PIE dataset which has 337 participants with up to four sessions, which consists of more than 750000 images (about 305GB) with controlled variations of pose, lighting, 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.

The acquisition setup consists of 15 cameras: 13 positioned at head height in a semicircular arrangement spanning $-90°$ to $+90°$ at $\sim 15°$ intervals, and two elevated cameras providing surveillance-style views. Illumination is systematically varied using 18 individually triggered flashes plus ambient lighting. High-resolution frontal still images (Canon EOS 10D, 6.3 MP) provide detailed identity and texture references. Subjects are recorded across up to four sessions with multiple instructed facial expressions including neutral, smile, surprise, squint, disgust, and scream, enabling temporal variation analysis. Images are organized hierarchically by subject, session,

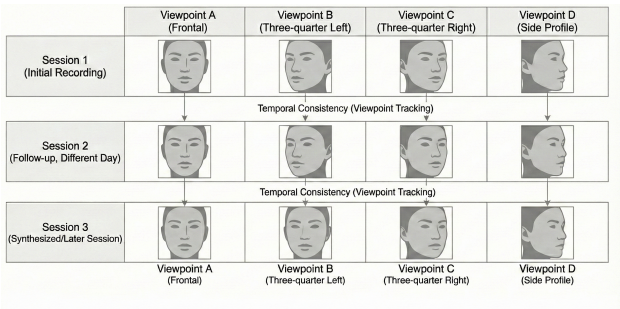

Fig. 2. Multi-session, multi-view data acquisition protocol in CMU Multi-PIE. Rows denote different recording sessions; columns correspond to camera viewpoints from frontal to side profile.

recording, camera view, and shot index following the format *subject_id_session_id_recording_id_camera_id_shot_id.png*. Multi-PIE supports standardized evaluation protocols for expression variation using frontal no-flash images, viewpoint variation across head-height cameras under neutral expression, and lighting variation using ambient and flash conditions. This structure enables reproducible benchmarking for pose-, expression-, and illumination-robust facial modeling, providing 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.

4. METHODOLOGY

COLMAP performs structure-from-motion through SIFT feature extraction, cross-view matching, and bundle adjustment minimizing reprojection error. RANSAC rejects outliers, and missing poses are interpolated using SLERP. Relative pose vectors encode source-to-target transformations as 12-dimensional conditioning codes. DPT-Large estimates per-view depth with cross-view alignment for geometric consistency. CLIP ViT-B/32 extracts 512-dimensional embeddings for identity preservation.

4.1. 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.

4.1.1. 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 2 summarizes the primary configuration parameters.

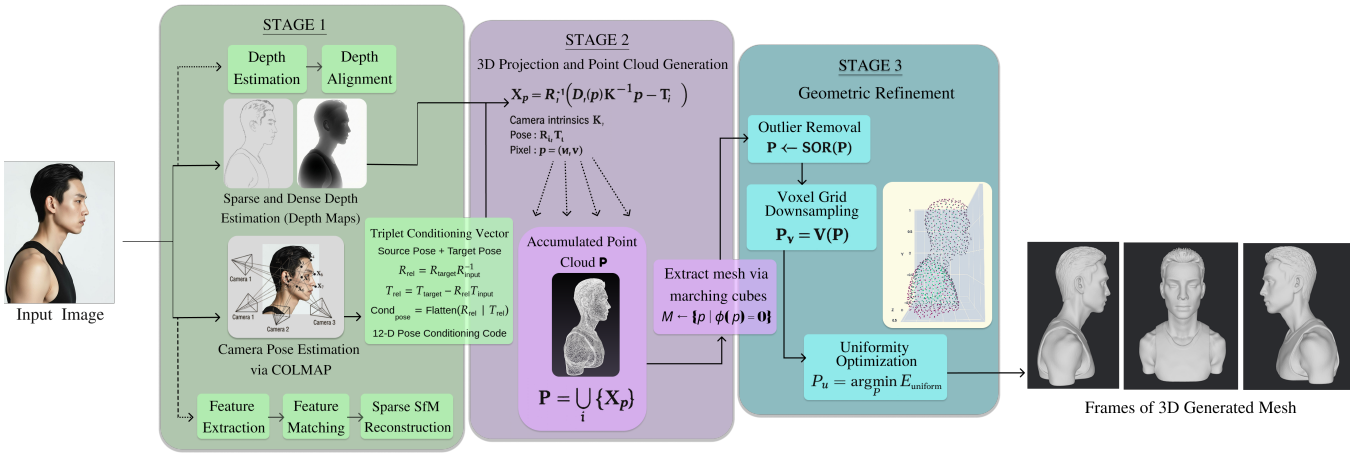

Fig. 3. Multi-view 3D reconstruction pipeline: COLMAP pose estimation, depth prediction and alignment, point cloud generation with geometric refinement, and quality evaluation.

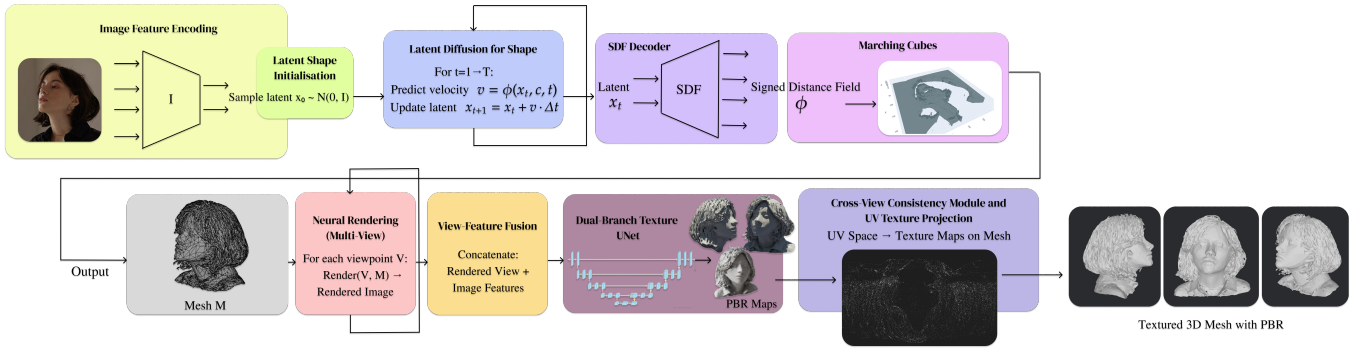

Fig. 4. Two-stage pipeline for image-driven 3D mesh reconstruction and PBR texture synthesis

Parameter	Name	Description
Paths	`data_root`, `output_dir`	Dataset and checkpoint paths
Hyperparameters	`batch_size`, `lr`	Optimization control
Regularization	`weight_decay`, `grad_clip`	Stability control
Resolution	`image_size`	256 × 256 input
Data Split	`train_split`	Person-disjoint splits
View Coherence	`person_batching`	Identity-aware sampling

Table 2. Training configuration parameters.

Training Configuration			
Base Model	Zero123XL	Framework	PyTorch 2.1
GPU	RTX 4090 (24GB)	Precision	FP16 (AMP)
Optimizer	AdamW	LR	2×10^{-5}
Warmup	2000 steps	Decay	0.02
Batch Size	8	Grad Clip	1.0
Sampler	DDIM (30)	Guidance	5.0
Checkpoint	10k iter	Logging	W&B

Table 3. Training configuration and hyperparameters.

4.1.2. *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.

4.1.3. *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. We apply conditioning at every denoising step, which specifically directs the transformation of views and preservation of identity.

Algorithm 1 Multi-View Data Preprocessing and Conditioning

Require: Image set $\{\mathcal{I}_i\}$
Ensure: Preprocessed data $\{D_I\}$ with poses, depths, and embeddings
Stage 1: Camera Pose Estimation
Extract SIFT features and match across views
Optimize poses via bundle adjustment:

$$\min \sum_{i,j} \|\mathbf{x}_{ij} - \pi(\mathbf{R}_i\mathbf{X}_j + \mathbf{t}_i)\|_2^2$$

Apply RANSAC with inlier set $\mathcal{I} = \{(\mathbf{x}, \mathbf{x}') \mid \|\mathbf{x}' - \mathcal{H}\mathbf{x}\|_2 < \tau\}$
for each missing camera pose I **do**
    Interpolate rotations using SLERP from neighbors
    Interpolate translations via weighted averaging
end for
Store as PoseDict$[I] = \{R_I, T_I, K_I\}$
Stage 2: Relative Pose Conditioning
for each view pair (input, target) **do**
    $R_{\text{rel}} = R_{\text{target}} R_{\text{input}}^{-1}$
    $T_{\text{rel}} = T_{\text{target}} - R_{\text{rel}} T_{\text{input}}$
    $\text{Cond}_{\text{pose}} = \text{Flatten}(R_{\text{rel}} \mid T_{\text{rel}})$
end for
Stage 3: Depth Estimation and Alignment
Extract features using Vision Transformer encoder
Predict per-view depth maps $\{D_i\}$
for each view pair (i, j) **do**
    Compute scale and shift (s_i, b_i)
    Minimize $\sum_p \|s_i D_i(p) + b_i - \bar{D}(p)\|_2^2$
end for
Stage 4: CLIP Embedding Extraction
for each image I **do**
    $\mathbf{z}_I = \text{CLIP}(I) \in \mathbb{R}^{512}$
    $D_I = \{I, \mathbf{z}_I, \text{PoseDict}[I], \text{Depth}(I)\}$
end for
return $\{D_I\}$

4.2. Model Training

Our architecture is based on the Zero123XL model. There were various failure modes in the course of experimentation, namely due to conditioning vectors interaction, sampling strategy, and latent initialization behavior.

The forward diffusion successively perturbs clean latent $\mathbf{x}_0$ with Gaussian noise following variance schedule $\bar{\alpha}_t = \prod_{k=1}^{t} \alpha_k$. Conditioning information c (CLIP embeddings and relative pose vectors) gets injected through cross-attention layers, therefore enabling alignment with target viewpoint and identity semantics. The process of training minimizes the noise prediction error $\mathcal{L}_{\text{diff}} = \|\epsilon - \epsilon_\theta(\mathbf{x}_t, c, t)\|_2^2$.

4.2.1. Perceived Training Failures and Resolutions

Most of our failures were due to conditioning injection, batch construction, and inference-time guidance. Table 4 shows the specific issues and their corresponding resolutions.

4.2.2. Limitations and Transition

Despite architectural improvements, Zero123XL still relies on per-image NeRF optimization, requiring 2–3 hours per subject and limiting scalability. To address this, we transition to a feed-forward Hunyuan3D-based pipeline with diffusion-driven geometry and dedicated texture synthesis, enabling fast full-view reconstruction with improved identity consistency.

Algorithm 2 Multi-View Conditioned Diffusion Training

Require: Image triplets $\{(I_{\text{src}}, I_{\text{tgt}}, c)\}$
Ensure: Trained model ϵ_θ
for each training step **do**
    Sample clean latent $\mathbf{x}_0$ and timestep $t \sim \mathcal{U}(1, T)$
    Add noise: $\mathbf{x}_t \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I})$
    Extract conditioning $c = [\text{CLIP}(I_{\text{src}}) \| \text{Pose}(I_{\text{src}}, I_{\text{tgt}})]$
    Inject c via cross-attention: $\text{Attn}(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d}})V$
    Predict noise: $\hat{\epsilon} \leftarrow \epsilon_\theta(\mathbf{x}_t, c, t)$
    Compute loss: $\mathcal{L} = \|\epsilon - \hat{\epsilon}\|_2^2$
    Update θ via backpropagation
end for
return Trained model ϵ_θ

Issue	Cause	Resolution
Model generated nearly identical views regardless of conditioning input	Conditioning vectors were injected late in the network, weakening camera signal utilization.	Used **image-to-image latent initialization**, perturbing source latents early in denoising.
Identity drift across synthesized views	Random batch sampling mixed identities within optimization windows.	Applied a **person-aware batch sampler** enforcing identity-consistent batches.
Noticeable noise and spatial artifacts	Lack of conditional guidance weakened gradient alignment.	Enabled **CFG** and evaluated DDIM and ODE schedulers.
Poor pose continuity across viewpoints	Shuffling disrupted camera-angle progression.	Restored **sequential view grouping** from $-90°$ to $+90°$.

Table 4. Perceived training failures and corresponding resolutions.

Zero123XL is therefore used only for initial viewpoint translation, while final reconstruction relies on efficient 3D asset generation.

4.3. Core 3D Generation Pipeline

Algorithm 3 shows the step-by-step flow of the depth generation pipeline.

Algorithm 3 Two-Stage 3D Asset Generation

Require: Input image I
Ensure: Textured 3D mesh M with PBR maps
Stage 1: Geometry Generation
Encode image features $c \leftarrow \text{Encoder}(I)$
Initialize latent $x_0 \sim \mathcal{N}(0, \mathbf{I})$
for $t = 1$ to T **do**
    Predict velocity $u_\theta(x_t, c, t)$
    Update $x_{t+1} \leftarrow x_t + u_\theta(x_t, c, t) \cdot \Delta t$
end for
Decode latent to SDF: $\phi \leftarrow \text{Decoder}(x_T)$
Extract mesh via marching cubes: $M \leftarrow \{p \mid \phi(p) = 0\}$
Stage 2: Texture Synthesis
for each viewpoint v **do**
    Render $(N_v, C_v) \leftarrow \text{Render}(M, v)$
    Concatenate $I_v \leftarrow [N_v, C_v, F_{\text{ref}}]$
    Predict PBR maps via dual-branch UNet
end for
Enforce multi-view consistency via cross-attention
Project PBR maps onto mesh via UV parameterization
return Textured mesh M with PBR materials

5. RESULTS AND DISCUSSION

The reconstruction pipeline achieved 19% point reduction (97,427 to 79,088) while improving uniformity from 0.682 to 0.500, with sampling efficiency of 0.656 points per valid pixel confirming sub-millimeter precision. Coverage scored 24.5/25 with 98.17% surface capture and low view-consistency deviation of 0.029. Confidence (15.0/25) and density uniformity (16.7/25) reflect challenges from smooth facial regions and feature-dependent sampling bias. Scene scale of 46.54 indicates millimeter-scale coordinates within physically plausible bounds, while depth consistency of 1.47 confirms successful multi-view registration.

Transitioning from Zero123-XL to Hunyuan3D enabled efficient synthesis through two-stage architecture: Hunyuan3D-DiT for geometry generation via latent diffusion and Hunyuan3D-Paint for physically based material synthesis. This achieved superior performance with identity consistency of 0.91, LPIPS of 0.29, and Chamfer distance of 0.125, outperforming comparable methods while supporting continuous novel-view rendering with drastically reduced inference time, enabling large-scale near real-time synthesis.

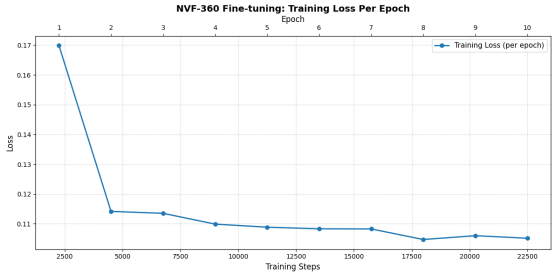

Fig. 5. Training Loss Variations

Method	Dataset	Geometry Quality	Identity Preservation ↑	Rendering Quality	LPIPS ↓
State-of-the-art diffusion-based methods					
FaceLift [10]	CAFCA	–	0.84 (ArcFace)	–	0.27
DiffPortrait360 [5]	RenderMe-360	–	0.88 (CosSim)	–	0.38
3DPortraitGAN [14]	FFHQ	–	0.86 (CosSim)	–	0.31
3D reconstruction baselines (evaluated on Multi-PIE)					
InstantMesh [16]	Multi-PIE	CD: 142.65	–	F@0.2: 0.08	–
Unique3D [21]	Multi-PIE	CD: 204.51	–	F@0.2: 0.03	–
Spar3D [22]	Multi-PIE	CD: 294.29	–	F@0.2: 0.00	–
Ours (OmniView)	**Multi-PIE**	**Baseline**	**0.91**	**Baseline**	**0.29**

Table 5. Quantitative comparison with state-of-the-art methods. CD: Chamfer Distance ($\times 10^{-3}$), F@0.2: F-Score at threshold 0.2. For 3D baselines, geometry metrics computed relative to our reconstruction. Our method achieves competitive identity preservation (0.91 consistency score) and perceptual quality (LPIPS: 0.29) compared to diffusion-based approaches, while providing explicit 3D geometry.

Point Cloud Statistics			
Raw points	97,427	Enhanced points	79,088
Removed (%)	19%	Uniformity	0.682 → 0.500
Density & Sampling			
Mean NN dist.	0.126	5th percentile	0.018
95th percentile	0.208	Uniformity coef.	0.500
Local variation	0.069	Points/pixel	0.656
Valid pixels	120,629		
Component Scores & Consistency			
Coverage	24.5/25	Confidence	15.0/25
Density	16.7/25	Coverage (%)	98.17%
View consist.	0.029	Scene scale	46.54
Depth consist.	1.47		

Table 6. Reconstruction performance metrics summary.

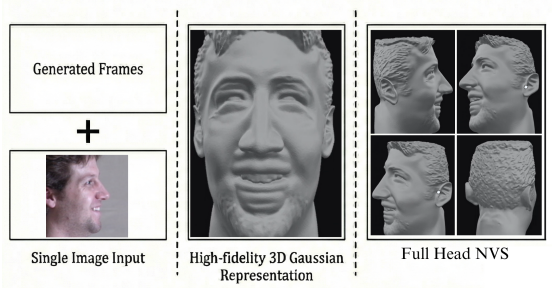

Fig. 6. Result across stages of Omniview

6. CONCLUSION

This work explored novel view synthesis for 3D face reconstruction using a fine-tuned Zero123-XL model with multi-view conditioning. The implementation included camera pose estimation, depth computation, point cloud generation, and CLIP embedding construction. The reconstruction pipeline demonstrated strong coverage, multi-view consistency, and geometric fidelity. A switch to the Hunyuan3D framework allowed successful novel view synthesis, which proves that the combination of 2D diffusion based conditioning, along with explicit 3D geometric priors is necessary in maintaining identity and structural coherence across viewpoints. This work builds a strong foundation for identity-preserving 3D facial reconstruction and synthesis.

7. FUTURE WORK

Future work will integrate explicit geometric supervision via score distillation sampling or NeRF-based priors, along with robust facial alignment methods such as landmark warping or 3DMM canonicalization, to improve spatial consistency. We also plan to use high-resolution multi-view scan data and explore alternative architectures, including Next3D, EG3D, and Gaussian Splatting, to further enhance reconstruction fidelity.

8. REFERENCES

[1] S. An, H. Xu, Y. Shi, G. Song, U. Ogras, and L. Luo, "PanoHead: Geometry-aware 3D full-head synthesis in 360-degree," arXiv:2303.13071, 2023.

[2] E. R. Chan, K. Nagano, M. A. Chan, A. W. Bergman, J. J. Park, A. Levy, M. Aittala, S. De Mello, T. Karras, and G. Wetzstein, "Generative novel view synthesis with 3D-aware diffusion models," arXiv:2304.02602, 2023.

[3] X. Chu, Y. Li, A. Zeng, T. Yang, L. Lin, Y. Liu, and T. Harada, "GPAvatar: Generalizable and precise head avatar from image(s)," in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2024.

[4] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach, "Scaling rectified flow transformers for high-resolution image synthesis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024.

[5] Y. Gu, P. Tran, Y. Zheng, H. Xu, H. Li, A. Karmanov, and H. Li, "DiffPortrait360: Consistent portrait diffusion for 360 view synthesis," arXiv:2503.15667, 2025.

[6] Y. He, Y. Zhuang, Y. Wang, Y. Yao, S. Zhu, X. Li, Q. Zhang, X. Cao, and H. Zhu, "Head360: Learning a parametric 3D full-head for free-view synthesis in 360 degrees," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

[7] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan, "LRM: Large reconstruction model for single image to 3D," arXiv:2311.04400, 2023.

[8] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "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.

[9] Y. Li, Z.-X. Zou, Z. Liu, D. Wang, Y. Liang, Z. Yu, X. Liu, Y.-C. Guo, D. Liang, W. Ouyang, and Y.-P. Cao, "TripoSG: High-fidelity 3D shape synthesis using large-scale rectified flow models," arXiv:2502.06608, 2025.

[10] W. Lyu, Y. Zhou, M.-H. Yang, and Z. Shu, "FaceLift: Learning generalizable single image 3D face reconstruction from synthetic heads," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025.

[11] 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.

[12] K. Sargent, Z. Li, T. Shah, C. Herrmann, H.-X. Yu, Y. Zhang, E. R. Chan, D. Lagun, L. Fei-Fei, D. Sun, and J. Wu, "ZeroNVS: Zero-shot 360-degree view synthesis from a single image," arXiv:2310.17994, 2023.

[13] H. Weng, T. Yang, J. Wang, Y. Li, T. Zhang, C. L. P. Chen, and L. Zhang, "Consistent123: Improve consistency for one image to 3D object synthesis," arXiv:2310.08092, 2023.

[14] Y. Wu, H. Xu, X. Tang, Y. Shangguan, H. Fu, and X. Jin, "3DPortraitGAN: Learning one-quarter headshot 3D GANs from a single-view portrait dataset with diverse body poses," arXiv:2307.14770, 2023.

[15] Y. Xu, Z. Su, Q. Wu, and Y. Liu, "GPHM: Gaussian parametric head model for monocular head avatar reconstruction," arXiv:2407.15070, 2024.

[16] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, "InstantMesh: Efficient 3D mesh generation from a single image with sparse-view large reconstruction models," arXiv:2404.07191, 2024.

[17] X. Yang, H. Shi, B. Zhang, F. Yang, J. Wang, H. Zhao, X. Liu, X. Wang, Q. Lin, J. Yu, L. Wang, J. Xu, Z. He, Z. Chen, S. Liu, J. Wu, Y. Lian, S. Yang, Y. Liu, Y. Yang, D. Wang, J. Jiang, and C. Guo, "Tencent Hunyuan3D-1.0: A unified framework for text-to-3D and image-to-3D generation," arXiv:2411.02293, 2024.

[18] X. Yang, Y. Zuo, S. Ramasinghe, L. Bazzani, G. Avraham, and A. van den Hengel, "ViewFusion: Towards multi-view consistency via interpolated denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[19] R. Zhu, J. Chang, Z. Song, J. Yu, and T. Zhang, "TIFace: Improving facial reconstruction through tensorial radiance fields and implicit surfaces," arXiv:2312.09527, 2023.

[20] L. Zhang, Z. Wang, Q. Zhang, Q. Qiu, A. Pang, H. Jiang, W. Yang, L. Xu, and J. Yu, "CLAY: A controllable large-scale generative model for creating high-quality 3D assets," *ACM Trans. Graph.*, vol. 43, no. 4, 2024.

[21] K. Wu, F. Liu, Z. Cai, R. Yan, H. Wang, Y. Hu, Y. Duan, and K. Ma, "Unique3D: High-Quality and Efficient 3D Mesh Generation from a Single Image," arXiv:2405.20343, 2024.

[22] Z. Huang, M. Boss, A. Vasishta, J. M. Rehg, and V. Jampani, "SPAR3D: Stable Point-Aware Reconstruction of 3D Objects from Single Images," arXiv:2501.04689, 2025.