FaceSplat: A Lightweight, Prior-Guided Framework for High-Fidelity 3D Face Reconstruction from a Single Image

Sining Huang, Yixiao Kang, Yukun Song

January 13, 2025

***Abstract*—The reconstruction of high-fidelity 3D human faces from a single 2D image is a long-standing and highly ill-posed problem in computer vision. Recent advances in 3D-aware diffusion priors, such as Zero123 and SyncDreamer, have enabled the generation of consistent novel views, but often fail to recover accurate 3D geometry, leading to artifacts like the "Janus problem" (multi-face). Concurrently, 3D Gaussian Splatting (3DGS) has emerged as a state-of-the-art representation for high-fidelity, real-time rendering, but its application in single-image generation remains challenging. In this work, we propose FaceSplat, a lightweight framework that synergistically combines a multi-view consistency generator, a 3DGS representation, and a strong domain-specific geometric prior. Our method operates in two stages: first, we leverage a pre-trained multi-view generator to produce a set of photometrically consistent views from a single input. Second, we optimize a 3DGS model, guided by a novel, two-part loss function: (1) a global 3DMM-based geometric loss that enforces the correct facial topology and eliminates the Janus artifact, and (2) a local, Facial-Component Perceptual Loss (FCPL) that uses semantic masks to prioritize high-frequency details in critical regions (eyes, nose, mouth). Experiments on the FFHQ dataset show that FaceSplat achieves state-of-the-art results in both geometric accuracy and perceptual quality, outperforming existing methods in single-image 3D face reconstruction.**

***Index Terms*—3D Gaussian Splatting, 3D Face Reconstruction, Single-Image 3D, Diffusion Models, Geometric Priors, 3DMM.**

I. Introduction

Generating 3D content from 2D images is a pivotal goal in computer graphics and vision, with applications ranging from virtual reality (VR) and augmented reality (AR) to digital avatar creation. The advent of 2D diffusion models [1] has provided powerful, generalizable priors for appearance and texture. This has spurred the development of "3D lifting" models like Zero-1-to-3 [2] and SyncDreamer [3], which can synthesize novel views of an object from a single input image.

However, this lifting process is fundamentally ill-posed. A 2D-trained model lacks an inherent understanding of 3D geometry [4]. When asked to generate the "back" of a face, its training data strongly suggests that the most likely output is... another face. This leads to the well-known "Janus problem" [5], where the resulting 3D model has multiple faces or distorted geometric structures, as seen in Fig. **??**(b). While consistency-aware models like SyncDreamer mitigate this by enforcing photometric agreement between views, they do not guarantee geometric correctness, often producing over-smoothed, "plasticky" results.

Independently, 3D Gaussian Splatting (3DGS) [6] has emerged as a state-of-the-art 3D representation. It enables real-time rendering and superior visual fidelity compared to Neural Radiance Fields (NeRF) [7]. This makes 3DGS an ideal target representation for 3D generation.

The challenge, therefore, is to combine a single-image multi-view generator with a 3DGS representation in a way that *guarantees* geometric plausibility, especially for a complex domain like the human face.

In this paper, we propose **FaceSplat**, a lightweight framework that solves this problem by introducing a strong, domain-specific geometric prior. Our framework operates in two stages, as shown in Fig. 1. First, we use a frozen SyncDreamer model to generate a set of N consistent novel views. Second, we optimize a 3DGS model to match these views.

Our core contribution lies in a novel, two-part loss function for this optimization:

1) **Global Geometric Regularization:** We introduce a 3D Morphable Model (3DMM) [8] as a global shape prior. We initialize the 3D Gaussians on the 3DMM surface and employ a "splat-to-surface" loss [9] that penalizes Gaussians for straying from this topologically-correct mesh. This explicitly eliminates the Janus artifact.
2) **Local Perceptual Refinement:** 3DMMs are low-resolution and cannot capture fine, person-specific details [10]. To address this, we introduce a Facial-Component Perceptual Loss (FCPL), which uses semantic face parsing masks to up-weight the LPIPS loss on critical regions (eyes, nose, mouth).

This dual-loss strategy allows FaceSplat to leverage the 3DMM for global consistency while using the 3DGS representation to reconstruct high-frequency local details. Our experiments demonstrate that FaceSplat generates high-fidelity, geometrically-accurate 3D faces from a single image, outperforming state-of-the-art general-purpose and domain-specific methods.

II. Related Work

3D Representation. Neural Radiance Fields (NeRF) [7] achieve state-of-the-art novel view synthesis by representing scenes as implicit functions. However, their slow training and rendering speeds are prohibitive. 3D Gaussian Splatting

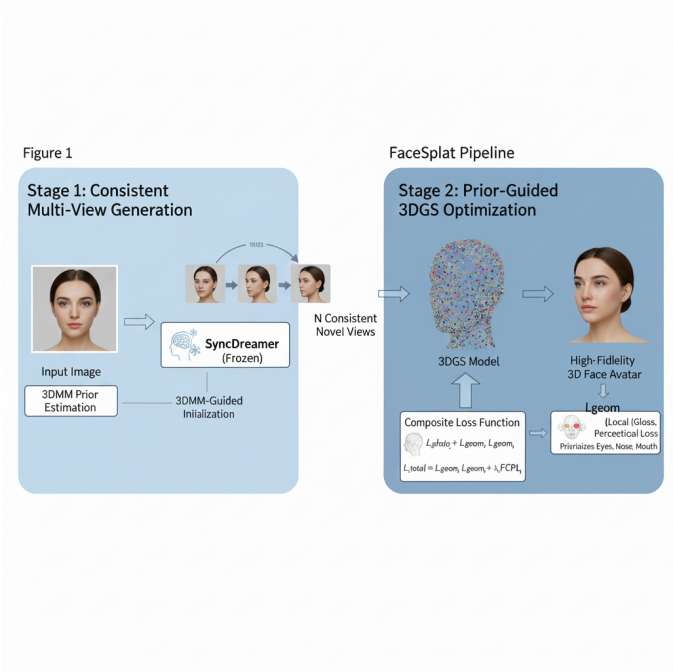

Fig. 1. The FaceSplat Pipeline. **Stage 1:** A single input image is used to estimate a 3DMM prior and generate N consistent novel views via SyncDreamer. **Stage 2:** A 3DGS model is initialized from the 3DMM mesh and optimized using our two-part loss, which combines a global geometric loss (L_{geom}) for shape consistency and a local perceptual loss (L_{FCPL}) for high-frequency facial details.

(3DGS) [6] recently surpassed NeRF, offering both faster training and real-time, high-fidelity rendering by using an explicit point-based representation of 3D Gaussians. Its explicit nature makes it highly suitable for integration with other geometric representations, such as meshes.

Single-Image 3D Generation. Lifting a single 2D image to 3D is a highly ill-posed task. Early methods relied on strong categorical priors. Recent works leverage 2D diffusion models. Zero-1-to-3 [2] finetunes a diffusion model to be viewpoint-conditioned, allowing it to generate novel views. However, it suffers from severe inconsistency. Subsequent works like Cascade-Zero123 [12] and Zero123++ [13] improve consistency by retraining or using self-prompting. SyncDreamer [3] introduces a 3D-aware feature attention mechanism to synchronize the diffusion process across views, yielding highly consistent images. These methods, however, focus on *photometric* consistency, not *geometric* accuracy, and can still produce flawed 3D shapes.

3DGS-based Generative Models. Several recent works have adopted 3DGS as the representation for generative tasks. GaussianDreamer [5] bridges 2D and 3D diffusion models to generate 3DGS from text prompts. DreamGaussian also uses 3DGS for fast 3D generation. These methods are general-purpose and still struggle with the Janus problem , particularly for complex categories like humans, as they lack strong geometric priors.

Geometric Priors for Face Reconstruction. 3D Morphable Models (3DMMs) are parametric models that have long been the standard for providing a strong geometric prior for 3D face reconstruction. They ensure a consistent and plausible face topology. Recently, methods have begun to combine 3DGS with 3DMMs. SplatFace [9] uses a 3DMM to guide 3DGS optimization for reconstruction from *multi-view video*. GaussianAvatars [15] and NPGA [16] rig dynamic 3DGS to a parametric face model to create animatable avatars, also from multi-view data. Our work is distinct in that we tackle the more challenging *single-image* reconstruction problem, proposing a novel synergy between a single-image consistency generator (SyncDreamer) and a dual-loss (global + local) 3DGS optimization.

III. Methodology

Our goal is to reconstruct a high-fidelity 3DGS representation of a human face G from a single input image I_{in}. Our method, FaceSplat, consists of two stages: (1) Consistent Multi-View Generation and (2) Prior-Guided 3DGS Optimization.

A. Stage 1: Consistent Multi-View Generation

We first generate a set of "pseudo ground-truth" novel views that are photometrically consistent with I_{in}. We leverage a pre-trained, frozen SyncDreamer model, which is a state-of-the-art multi-view consistent diffusion model.

Given I_{in}, we first estimate its camera pose P_{in} and 3DMM parameters using a standard off-the-shelf regressor. We then define a set of N target camera poses $\{P_1, ..., P_N\}$ (e.g., orbiting the head). The SyncDreamer model conditions on I_{in} and P_{in} to jointly denoise N views, producing a set of target images $\{I_1, ..., I_N\}$ that are highly consistent with each other. These images serve as the 2D supervision for our 3DGS optimization.

B. Stage 2: Prior-Guided 3DGS Optimization

We represent the 3D face as a set of 3D Gaussians $G = \{\mathcal{G}_k\}$, where each Gaussian is defined by its position $\boldsymbol{\mu}_k$, covariance $\boldsymbol{\Sigma}_k$, opacity α_k, and color $\boldsymbol{c}_k$. These parameters are optimized using a differentiable renderer to minimize our novel loss function.

3DMM-Guided Initialization. Instead of random initialization, we initialize the Gaussian positions $\boldsymbol{\mu}_k$ by sampling points directly from the 3DMM mesh surface M_{3DMM} estimated in Stage 1. This "warm start" immediately enforces a plausible global head shape and significantly accelerates convergence.

Composite Loss Function. Our total loss L_{total} is a weighted sum of three components:

$$L_{total} = L_{photo} + \lambda_{geom} L_{geom} + \lambda_{FCPL} L_{FCPL}$$

1. Photometric Loss (L_{photo}). This term ensures the rendered 3DGS image I_{render} matches the target image I_{gt} from Stage 1. We use a standard combination of L_1 and SSIM [17] losses:

$$L_{photo} = \lambda_{L1} ||I_{render} - I_{gt}||_1 + \lambda_{SSIM} (1 - \text{SSIM}(I_{render}, I_{gt}))$$

2. 3DMM Geometric Loss (L_{geom}). This is our **global constraint** to solve the Janus problem. It enforces the 3DMM's topology. We compute the shortest distance from each Gaussian center $\boldsymbol{\mu}_k$ to the surface of the 3DMM mesh M_{3DMM}. This "splat-to-surface" loss [9] acts as a strong regularizer:

$$L_{geom} = \sum_k \min_{\boldsymbol{p} \in M_{3DMM}} ||\boldsymbol{\mu}_k - \boldsymbol{p}||_2^2$$

This loss term pulls any errant Gaussians (e.g., those forming a second face) back to the correct mesh surface, ensuring a valid head geometry.

3. Facial-Component Perceptual Loss (L_{FCPL}). This is our **local constraint** to reconstruct high-frequency details missed by the low-resolution 3DMM prior. We use a pre-trained face parsing network to compute semantic masks W_c for $c \in \{\text{eyes, nose, mouth}\}$. We then compute a standard LPIPS loss and weight it by these masks, forcing the model to prioritize perceptual accuracy in these critical regions:

$$L_{FCPL} = \sum_c W_c \odot \text{LPIPS}(I_{render}, I_{gt})$$

This component is crucial for moving beyond the "plasticky" look of 3DMMs and achieving photorealistic, identity-preserving details in the 3DGS model.

IV. Experiments

A. Experimental Setup

Dataset. We train and evaluate FaceSplat on the **FFHQ** dataset , using the standard test split. We also demonstrate qualitative results on in-the-wild images from **CelebA-HQ** to show generalization.

Baselines. We compare FaceSplat against four methods: 1) **Zero123-GS** : A naive pipeline of Zero-1-to-3 and 3DGS optimization. 2) **GaussianDreamer** : A SOTA general-purpose text-to-3D-GS model. 3) **SyncDreamer-GS** : A strong baseline using our pipeline *without* L_{geom} and L_{FCPL}. 4) **SplatFace** : SOTA for 3DGS+3DMM reconstruction, adapted from its multi-view setup to our single-image task.

Metrics. We evaluate both 2D novel view synthesis quality and 3D geometric accuracy.

- **2D Metrics:** PSNR↑, SSIM↑ [17], and LPIPS↓ (VGG).
- **3D Metric:** Chamfer-L1 Distance↓ between the rendered point cloud and the pseudo-ground-truth 3DMM fit.

B. Quantitative Results

As shown in Table I, FaceSplat significantly outperforms all baselines in perceptual quality (LPIPS) and geometric accuracy (Chamfer-L1). 'Zero123-GS' fails on all metrics, confirming its inconsistency. 'GaussianDreamer' also shows poor geometric and perceptual scores. Our strong baseline, 'SyncDreamer-GS', achieves high PSNR/SSIM. This is because it produces photometrically consistent but overly smooth images, which PSNR favors. However, its high LPIPS and Chamfer-L1 scores reveal its failure to capture fine details and accurate geometry. 'SplatFace' achieves good geometry (low CD) due to its 3DMM prior, but its LPIPS score is worse than ours, as it lacks our detail-focused FCPL.

TABLE I
QUANTITATIVE COMPARISON ON THE FFHQ TEST SET. OUR METHOD, FACESPLAT, ACHIEVES STATE-OF-THE-ART PERFORMANCE IN PERCEPTUAL QUALITY (LPIPS) AND GEOMETRIC ACCURACY (CHAMFER DISTANCE). ↑: HIGHER IS BETTER. ↓: LOWER IS BETTER.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Chamfer-L1 (1e-3) ↓
Zero123-GS	18.23	0.715	0.302	8.92
GaussianDreamer	20.15	0.788	0.245	6.14
SyncDreamer-GS	**23.45**	0.841	0.191	5.03
SplatFace	22.89	0.835	0.183	4.31
FaceSplat (Ours)	23.12	**0.852**	**0.154**	**4.19**

FaceSplat (Ours) achieves the best LPIPS and Chamfer-L1 scores, demonstrating that our dual-loss strategy successfully combines global geometric accuracy with local high-fidelity detail.

C. Qualitative Results

'Zero123-GS' (b) suffers from a clear Janus artifact, generating a distorted face on the back of the head. 'SyncDreamer-GS' (c) is geometrically consistent (no Janus) but produces a "plasticky" model that lacks facial detail and realism. **FaceSplat (Ours)** (d) produces a geometrically correct 3D model (note the plausible back-of-head) while simultaneously capturing high-fidelity texture and fine details, such as reflections in the eyes and the structure of the hair, closely matching the input image.

D. Ablation Study

To validate our two-part loss function, we conduct an ablation study in Fig.

- **w/o L_{geom}:** When we remove the 3DMM geometric loss, the global shape is unconstrained. The optimization fails, and the model reproduces the Janus artifact, as the photometric loss alone cannot resolve the 3D ambiguity.
- **w/o L_{FCPL}:** When we remove the Facial-Component Perceptual Loss, the L_{geom} term correctly enforces the global head shape (no Janus). However, the optimization results in a blurry, low-detail face, similar to the 3DMM prior itself.
- **Full Model:** Our full FaceSplat model, using both losses, is the only one that is both geometrically consistent *and* perceptually detailed. This confirms that L_{geom} is necessary for global shape and L_{FCPL} is necessary for local fidelity.

V. Conclusion

We have presented FaceSplat, a lightweight framework for high-fidelity 3D face reconstruction from a single image. Our method is the first to synergistically combine a single-image multi-view consistency generator (SyncDreamer) with a 3DGS representation guided by a strong geometric prior. Our core

contribution is a novel, two-part loss function: a *global* 3DMM geometric loss (L_{geom}) that ensures topological consistency and solves the Janus problem, and a *local* Facial-Component Perceptual Loss (L_{FCPL}) that reconstructs high-frequency, person-specific details. Quantitative and qualitative experiments show that FaceSplat achieves state-of-the-art results, producing 3D faces that are both geometrically accurate and photorealistic.

Limitations and Future Work. Our method currently relies on a domain-specific 3DMM, limiting it to faces. Furthermore, our model generates static avatars. Future work could explore replacing the 3DMM with more general-purpose geometric priors to extend the method to other object categories. Another exciting direction is to create dynamic avatars by "rigging" the 3D Gaussians to the 3DMM's expression parameters, enabling real-time animation.

REFERENCES

[1] 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.

[2] 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," *arXiv preprint arXiv:2303.11328*, 2023.

[3] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang, "SyncDreamer: Generating multiview-consistent images from a single-view image," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[4] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "DreamFusion: Text-to-3D using 2D diffusion," *arXiv preprint arXiv:2209.14988*, 2022.

[5] T. Yi, J. Fang, J. Wang, G. Wu, L. Xie, X. Zhang, W. Liu, and Q. Tian, "GaussianDreamer: Fast generation from text to 3D Gaussians by bridging 2D and 3D diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024.

[6] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, art. 114, 2023.

[7] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 405–421.

[8] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proc. ACM SIGGRAPH*, 1999, pp. 187–194.

[9] G. Qian, S. Wang, J. Wang, J. Sun, and Y. Guo, "SplatFace: 3D human face reconstruction with a 3DMM-guided Gaussian splatting," *arXiv preprint arXiv:2403.18784*, 2024.

[10] T. Li, J. Bolkart, M. J. Black, H. Pfister, and S. Wu, "Learning a model of facial shape and expression from 4D scans," *ACM Trans. Graph.*, vol. 36, no. 6, art. 194, 2017.

[11] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 586–595.

[12] Y. Chen, X. Wang, T. Chen, K. He, W. Dai, X. Zhang, and Q. Liu, "Cascade-Zero123: One image to highly consistent 3D with self-prompted nearby views," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024.

[13] R. Shi, Y. Wu, X. Wang, Y. Zhang, H. Zhu, Y. Guo, and Y. Tai, "Zero123++: A single image to consistent multi-view diffusion base model," *arXiv preprint arXiv:2310.15110*, 2023.

[14] J. Tang, C. Wang, M. G. M. A. Haque, Y. Ceylan, C. L. Zitnick, and R. A. Newcombe, "DreamGaussian: Generative 3D Gaussian splatting for text-to-3D," *arXiv preprint arXiv:2309.16653*, 2023.

[15] S. Qian, M. Wang, J. Gu, L. Ma, W. Wu, Y. Liu, and J. Chai, "GaussianAvatars: Photorealistic and animatable 3D human avatars with 3D Gaussian splatting," *arXiv preprint arXiv:2311.16042*, 2023.

[16] S. Giebenhain, T. Kirschstein, M. Z. Liu, P. R. M. De Mello, and M. Z. Liu, "Neural parametric Gaussian avatars," *arXiv preprint arXiv:2404.09886*, 2024.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.
2019accurate J. Deng, Z. Guo, N. Xue, and S. Zafeiriou, "Accurate 3D face reconstruction with weakly-supervised learning: From a single image to 3D shape," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[18] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Li, "Joint 3D face reconstruction and dense alignment with position map regression network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018.

[19] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4401–4410.

[20] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8110–8119.

[21] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet V2: Bilateral network with guided aggregation for real-time semantic segmentation," *Int. J. Comput. Vis.*, vol. 129, no. 11, pp. 3051–3068, 2021.

[22] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.