Faithful Single Image Face Reconstruction Using GAN Inversion

Yiming Luo
Imperial College London
g8982988@gmail.com

Abhijeet Ghosh
Imperial College London
abhijeet.ghosh@imperial.ac.uk

Abstract

Three-dimensional generative adversarial networks (3D GANs) enable high-fidelity and view-consistent image synthesis for applications such as digital humans and avatars. However, existing inversion methods face key limitations. Encoder-based approaches, though efficient, often lose fine details and degrade under challenging poses. Optimization-based methods, while flexible, lack geometry supervision and produce distortions in novel views. Symmetry priors partly address these issues but restrict generalization and rely on fixed pose estimators. Moreover, pose-conditioned 3D GANs assume known pose distributions, which may misalign with real data and misguide geometry learning. To overcome these challenges, we propose an optimization-based 3D GAN inversion framework built upon a pose-free generator that implicitly learns the pose distribution. This removes the need for pose annotations or rigid priors, while reformulating inversion as a latent-only optimization problem. We further introduce geometry-aware regularization to stabilize latent optimization and improve view consistency. Experiments on challenging benchmarks demonstrate that our approach achieves more faithful reconstructions, geometrically consistent novel views, and robust identity preservation under extreme poses and asymmetric structures compared to prior methods.

Keywords: 3D GAN Inversion, Novel View Synthesis, Single View Face Reconstruction.

1. Introduction

Recent advances in three-dimensional generative adversarial networks (3D GANs) have enabled high-fidelity, view-consistent image synthesis from monocular inputs, supporting applications such as 3D avatar creation, facial reenactment, virtual reality, and digital content generation [4, 26, 14]. By modeling underlying geometry using neural implicit representations or hybrid volumetric renderers, 3D GANs such as EG3D [4] have become foundational for enabling explicit 3D control in image synthesis. To fully exploit these models, it is essential to address the task of 3D GAN inversion, which aims to recover a latent code that allows the generator to faithfully reconstruct the input image and enable view-consistent novel rendering, while supporting semantic editing in the latent space [25].

Figure 1. Given a single input image, our method reconstructs a photorealistic and identity-preserving 3D face, enabling consistent novel view synthesis.

Despite these advances, 3D GAN inversion remains challenging due to the ill-posed nature of monocular-to-3D reconstruction and the highly entangled latent space of GANs [25]. Existing methods can be broadly categorized into encoder-based and optimization-based approaches. Encoder-based methods, such as E3DGE [14] and IDE-3D [22], predict latent codes via feedforward encoders, enabling real-time inversion. However, as acknowledged in E3DGE [14], these methods struggle to preserve fine-grained details and stable geometry, particularly under challenging poses, and their reconstruction quality generally lags behind optimization-based approaches.

Optimization-based methods, including Pivotal Tuning Inversion (PTI) [18], originally designed for 2D inversion, improve fidelity by locally fine-tuning the generator. Yet, as noted by SPI [26], directly applying PTI to 3D GANs neglects geometry supervision, resulting in geometry collapse and severe distortions in novel views. SPI [26] mitigates these issues by introducing a facial symmetry prior to provide pseudo multi-view supervision. However, its reliance on symmetry assumptions and fixed camera poses limits generalization, particularly when handling asymmetric faces, occlusions, or side-view inputs, where symmetry-based supervision can introduce artifacts rather than suppress them.

In addition, Pose-Optimized inversion [13] introduces an explicit pose refinement branch within the inversion loop. While this improves flexibility compared to fixed-pose methods, its inversion accuracy still depends on the quality of the initial pose estimate produced by a learned pose regressor, making it sensitive to pose ambiguities and less robust in unconstrained settings.

Moreover, most existing 3D GAN inversion frameworks operate within pre-trained pose-conditioned models such as EG3D [4], assuming known pose distributions and disentangled geometry and appearance. As highlighted by Shi et al. [21, 20], these designs are fundamentally limited by their reliance on manually designed pose priors, which, when mismatched with real-world distributions, can mislead geometry learning and result in degenerate solutions.

To address these challenges, we propose an optimization-based 3D GAN inversion framework built upon a pose-free 3D generator. By leveraging the generator's internally learned viewpoint representation, our method reformulates inversion as optimization without requiring explicit pose estimation, symmetry assumptions, or predefined pose priors. This design simplifies the inversion process and enables more stable geometry reconstruction and identity preservation under diverse viewpoints. Our contributions are summarized as follows:

- We propose an optimization-based 3D GAN inversion framework built upon a pose-free generator, enabling faithful reconstruction and novel-view synthesis without explicit pose estimation or external pose priors.
- We introduce geometry-aware regularization and multi-view consistency constraints to stabilize latent optimization, improving identity preservation and 3D consistency without relying on popular design like symmetry priors.
- We conduct extensive experiments showing that our framework outperforms existing methods in reconstruction fidelity, geometric consistency, and identity preservation, particularly under extreme poses and asymmetric structures.

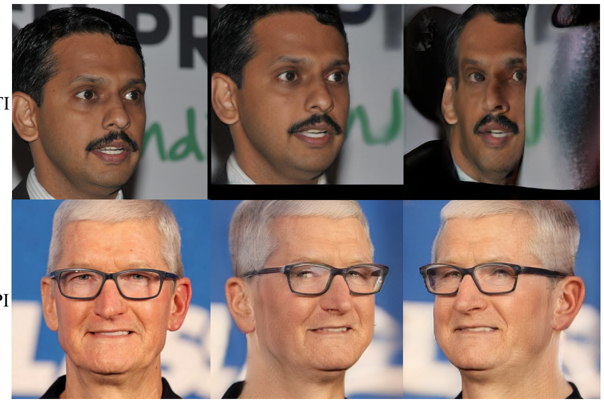

Figure 2. Failure cases of existing state-of-the-art optimization-based inversion methods.

2. Related Work

Single-image 3D face reconstruction. Prior to 3D-aware GANs, single-image facial reconstruction was primarily addressed through 3D Morphable Model (3DMM)-based approaches. Methods such as Deep3DFaceRecon [7] and DECA [8] fit parametric face models to monocular images via differentiable rendering, recovering geometry, expression, and pose parameters. While these methods provide explicit 3D representations, they are constrained by the limited expressiveness of parametric face spaces and cannot model appearance details such as hair, accessories, or background. Notably, many 3D GAN inversion methods rely on such 3DMM-based pose estimators as a prerequisite, inheriting their estimation errors. Our method avoids this dependency by building upon a pose-free generator that learns viewpoint distributions implicitly.

3D-aware GANs and Limitations The integration of 3D priors into GANs has enabled view-consistent image synthesis from monocular inputs. Early voxel-based methods, such as HoloGAN [15] and GRAF [19], were limited by low resolution and view inconsistency. Neural implicit representations, including pi-GAN [3] and GIRAFFE [16], enabled continuous geometry modeling but suffered from high computational cost and instability. EG3D [4] introduced a hybrid triplane-based renderer, improving both efficiency and quality, yet relying on manually designed pose priors or pre-estimated poses, limiting its robustness in unconstrained settings [20]. Recent works, such as PoF3D [20] and StyleNeRF [9], propose implicit pose learning within unconditional 3D GANs, removing explicit pose conditioning. However, these methods mainly focus on unconditional generation, leaving inversion underexplored.

2D GAN Inversion. GAN inversion has been extensively studied in 2D settings [25]. Encoder-based methods, such as pSp [17], e4e [23], ReStyle [2], and Style-Transformer [11], offer fast inversion via learned encoders but often sacrifice reconstruction quality and editability. Optimization-based methods, including Image2StyleGAN [1] and PTI [18], provide high-fidelity results through iterative latent refinement but suffer from inefficiency and overfitting risks.

3D GAN Inversion. Extending inversion into 3D GANs introduces additional challenges. IDE-3D [22] and E3DGE [14] propose encoder-based inversion for 3D GANs, achieving real-time inversion but struggling to preserve identity and geometry under extreme poses. SPI [26] introduces a facial symmetry prior to guide 3D inversion, but its reliance on symmetry assumptions and fixed poses makes it sensitive to asymmetric faces, occlusions, or side-view inputs, where it remains prone to artifacts despite conflict mask refinements. Pose-Optimized inversion [13] attempts to jointly refine pose and inversion, yet its inversion accuracy remains limited by the reliance on a simple pose regressor, making it sensitive to initial pose estimation errors and less robust in unconstrained scenarios.

3. Method

3.1. Preliminaries and Overview

GAN inversion aims to find a latent code of a pre-trained GAN such that the generator can reconstruct (and subsequently manipulate) a given input image. In the context of 3D GAN inversion, the goal is to associate an input image I with the latent space of a pre-trained 3D-aware generator G_{3D}, so that G_{3D} can produce a faithful reconstruction and generate view-consistent novel renderings under its internal 3D representation.

Traditional 3D GAN inversion methods often build upon camera-conditioned generators such as EG3D [4], which typically require camera parameters (or a pose prior) during training and provide camera conditioning during inversion. In practice, however, accurate camera annotations are rarely available for in-the-wild datasets. As a result, methods commonly rely on approximate pose priors (e.g., assuming yaw angles within $\pm 45°$ for face datasets) or off-the-shelf pose predictors. Both strategies can be fragile: inaccurate priors yield inconsistent geometry, while errors in pose prediction propagate directly into the inversion process. A common alternative is to jointly optimize the latent code and the camera pose. Although this removes the need for external pose labels, it expands the search space and introduces pose–appearance entanglement, often leading to unstable convergence and degraded identity preservation.

A representative attempt to operationalize this idea is Pose-Optimized inversion (Pose Opt) [13], which explicitly treats the camera pose as an optimization variable. While effective in principle, the joint optimization over latent codes and camera parameters enlarges the search space, and prior studies have noted that such strategies can be sensitive to pose initialization. Without a reasonable starting pose, optimization may converge to suboptimal solutions in which pose errors are absorbed by latent updates, potentially resulting in degraded geometry. Additional regularization terms, such as depth consistency or landmark constraints, are often introduced to stabilize the process, which increases complexity and computational cost.

Inspired by recent pose-free 3D-aware generators [4, 21, 20], we adopt a different perspective: rather than explicitly optimizing camera parameters, we reduce inversion to latent-only optimization and obtain an internally consistent viewpoint through the generator's learned pose prediction. Formally, our inversion solves

$$w^* = \arg\min_{w} \ \mathcal{L}\big(G_{3D}(w),\ I\big) \tag{1}$$

where the camera pose $\hat{\theta} = f_{\text{pose}}(w)$ is a deterministic function of the latent code. This design avoids reliance on external pose annotations or fragile pose initialization. To achieve high-quality reconstructions and view-consistent synthesis, we introduce four complementary losses organized in a coarse-to-fine hierarchy: single-view reconstruction objectives first ensure appearance fidelity, while multi-view consistency and geometric regularization then consolidate 3D structure.

3.2. Inversion Framework

Inspired by recent rapidly development of 3D GANs [4, 21, 20], our inversion framework builds upon recent advances in pose-free 3D-aware generative models [20] that decouple geometry learning from explicit camera supervision during training. Unlike conventional camera-conditioned GANs, where the generator requires externally supplied camera parameters as input, our design integrates pose prediction as an internal component of the generator. This enables inversion to be formulated as latent optimization only, while an internally consistent viewpoint is obtained implicitly during rendering. More details are in the supplementary files.

Latent-driven generator. At the core of the generator is a mapping from a latent code w to both an image and a corresponding camera pose. Specifically, a lightweight sub-network, referred to as the pose learner, maps the latent representation to a pose parameter $\hat{\theta}$, which is then used by a neural radiance field (NeRF)-based renderer to synthesize the final image. The pose learner and generator are trained jointly in an unsupervised manner, allowing the model to capture an internally consistent distribution of viewpoints directly from data, without requiring pose annotations. From the inversion perspective, this design makes the generator *latent-driven*: optimizing the latent code suffices to obtain a reconstruction whose 3D representation and associated internal viewpoint are consistent under the generator's canonical frame.

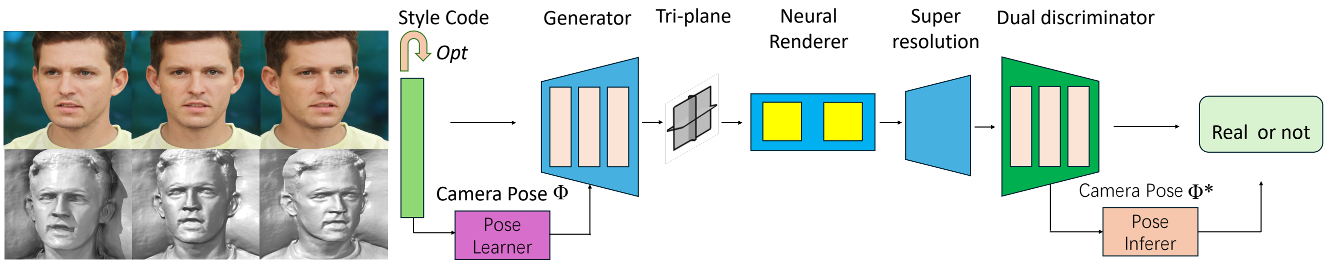

Figure 3. Overview of our pose-free 3D GAN inversion framework. Our method performs optimization-based inversion within a pose-free 3D GAN generator, enabling geometry-consistent reconstruction and novel-view synthesis from a single image without explicit camera pose estimation.

Pose–appearance consistency discriminator. To regularize pose learning during generator training, the framework employs a discriminator that conditions on both the rendered image and an associated pose signal [21, 20]. Concretely, the discriminator operates on image–pose pairs (I, θ), where generated samples are paired with the pose predicted by the pose learner, and real images are paired with pose proxies inferred by the discriminator's pose-prediction branch. This discriminator encourages consistency between pose signals and pose-dependent appearance cues, discouraging degenerate pose–appearance coupling (e.g., pose collapse or encoding pose into texture) during adversarial training, leading to more stable geometry learning even without explicit pose or depth supervision.

Adversarial training scheme. The generator, pose learner, and discriminator are trained end-to-end in an adversarial manner. During training, latent codes are sampled from a prior distribution and mapped to images by the generator, with the pose learner predicting the corresponding pose parameters. The discriminator distinguishes real image–pose pairs from generated ones, using pose proxies for real images inferred by its pose-prediction branch. Through this adversarial interaction, the generator learns a pose-consistent internal representation that does not rely on manually specified pose priors, such as fixed yaw or pitch ranges, which are often inaccurate for in-the-wild datasets [20].

Inversion perspective. After training, the generator and pose learner are fixed, and inversion is performed by optimizing only the latent code. In contrast to pipelines that require external pose predictors or joint optimization over latent and pose variables, the pose associated with a reconstructed image is obtained implicitly as a function of the optimized latent code. While this does not recover the true camera pose in an absolute sense, it yields an internally consistent viewpoint under the generator's learned canonical frame. Removing explicit pose variables simplifies the inversion procedure and, as we demonstrate in Section 4, improves robustness to challenging viewpoints and unaligned inputs.

To summarize, we adopt the pose-free 3D-aware generator of [4, 21, 20], which provides adversarial pose–appearance regularization, and perform optimization-based inversion on the frozen generator. By reframing inversion as a latent-only optimization problem, we avoid explicit pose supervision while preserving 3D consistency, setting the stage for the loss design described next.

3.3. Loss Design

Reducing inversion to latent-only optimization removes explicit pose variables, but it also requires carefully designed objectives to avoid degenerate solutions and encourage stable 3D structure. We introduce four complementary losses that address distinct aspects of inversion: reconstruction fidelity, identity or semantic preservation, multi-view consistency, and geometric regularization. Together, these objectives provide effective constraints without relying on explicit camera supervision.

Reconstruction loss. The primary objective is to align the rendered image $\hat{I}$ with the input image I. We combine pixel-wise and perceptual terms:

$$\mathcal{L}_{\text{rec}} = \alpha \|I - \hat{I}\|_1 + \beta \, \text{LPIPS}(I, \hat{I}) \tag{2}$$

where LPIPS [27] encourages perceptual similarity in deep feature space. The pixel loss preserves low-frequency structure such as global shape and color, while LPIPS promotes perceptual fidelity and discourages overly smooth reconstructions.

Identity loss. For portrait images, identity preservation is critical. We employ a face recognition network, ArcFace [5], and minimize the cosine distance between embeddings:

$$\mathcal{L}_{\text{id}} = 1 - \cos(\phi(I), \phi(\hat{I})) \tag{3}$$

where $\phi(\cdot)$ denotes the embedding function. For non-face domains, where identity-specific metrics are unavailable, we instead use CLIP [**?**] image features as a semantic consistency constraint. In this case, $\mathcal{L}_{\text{id}}$ encourages preservation of global semantic content rather than strict identity.

Multi-view consistency loss. To regularize 3D structure, we encourage local consistency across nearby viewpoints. Given a latent code w, we perturb its predicted pose $\hat{\theta}(w)$ by small offsets $\Delta\theta$ and render corresponding novel views $\hat{I}_{\Delta\theta}$. Specifically, we keep the latent code fixed and directly modify the pose input of the renderer using $\hat{\theta}(w) + \Delta\theta$. Using the depth map rendered by the generator, we warp these views back to the reference viewpoint, producing $\hat{I}_{\Delta\theta\to 0}$. We then define:

$$\mathcal{L}_{\text{mv}} = \sum_{\Delta\theta} \left\| \psi\left(\hat{I}_{\Delta\theta\to 0}\right) - \psi\left(\hat{I}\right) \right\|_1 \tag{4}$$

where $\psi(\cdot)$ denotes perceptual features. This loss encourages locally consistent appearance under small viewpoint changes within the generator's learned pose manifold, discouraging floating textures and view-dependent artifacts.

Geometric regularization. We further impose a weak geometric prior to suppress spurious structures in the implicit representation. Specifically, we apply a total variation penalty on the rendered depth map D:

$$\mathcal{L}_{\text{geo}} = \text{TV}(D) \tag{5}$$

This term encourages smooth depth variation and suppresses high-frequency noise in the rendered geometry without enforcing explicit shape supervision.

Overall objective. The full inversion objective is a weighted combination of the above losses:

$$\mathcal{L} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} + \lambda_{\text{mv}}\mathcal{L}_{\text{mv}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}} \tag{6}$$

Each term addresses a specific failure mode of inversion, and together they provide a stable and effective optimization objective. More details are in the supplementary file.

3.4. Optimization Strategy

We formulate inversion as an optimization over the latent code w with respect to the objective in Eq. (6), while keeping the generator and pose learner fixed.

To improve stability, we adopt a two-stage optimization strategy. In a pose-free generator, the latent code simultaneously encodes both appearance and an implicit viewpoint, making the optimization landscape more complex than in camera-conditioned settings where pose is fixed externally. To address this, the first stage performs optimization at a lower spatial resolution (256^2) to align global structure and the associated internal viewpoint without being distracted by high-frequency details. In the second stage, we switch to the native resolution of the generator (512^2) to refine fine-grained content while gradually strengthening geometry-related regularization.

We optimize the latent code using the Adam optimizer with cosine learning-rate decay. The relative weights of the losses are adjusted across stages following a curriculum schedule: in early iterations, reconstruction and identity terms dominate, ensuring faithful alignment with the input image. As optimization proceeds, we reduce the emphasis on identity preservation while increasing the strength of multi-view consistency and geometric regularization. This curriculum-style scheduling encourages the optimization to first match appearance and then consolidate 3D stability. Specific hyperparameter values and scheduling details are provided in the supplementary material.

As a result, the two-stage strategy improves convergence stability and yields reconstructions that remain consistent across viewpoints.

4. Experiments

4.1. Overview

In this section, we conduct several experiments to validate the effectiveness of our proposed method. We evaluate our approach on the CelebA-HQ [12] and demonstrate its capability in terms of reconstruction quality and identity preservation. Our method is compared against state-of-the-art approaches including SPI [26], pSp [17], PTI [18], 3D GAN Inversion with Pose Optimization (Pose Opt) [13], E3DGE [14], IDE-3D [22], and e4e [23]. All experiments are conducted on an NVIDIA RTX 4090 GPU. For quantitative evaluation, we employ commonly used metrics, including MSE, LPIPS [27], ID Similarity, and MS-SSIM [24], which respectively assess reconstruction accuracy, perceptual similarity, identity preservation, and structural quality.

4.2. Comparison with GAN-Inversion Methods

We first evaluate the reconstruction quality of our method on the CelebA-HQ [12] dataset. evaluating our method against seven state-of-the-art inversion methods, including SPI [26], pSp [17], PTI [18], Pose Opt [13], E3DGE [14], IDE-3D [22], and e4e [23]. To ensure fairness, all output images are resized to 256×256 in the experiments conducted in this section. As illustrated in Fig. 4, while all methods produce plausible results, they exhibit varying degrees of artifacts or identity inconsistencies. Specifically, Pose Opt [13] often suffers from inaccuracies introduced by its learned pose estimator, leading to noticeable misalignment and degraded reconstruction quality. PTI [18], although capable of generating high-fidelity results, struggles to consistently preserve the input identity.

In contrast, our method leverages implicit pose distribution learning to enhance robustness to pose variations and faithfully preserve intricate facial details. As is also shown in Fig. 5, our method reconstructs more accurate and geometry-consistent 3D facial structures from a single image, effectively maintaining identity coherence even under challenging viewpoints.

Quantitative results in Table 1 further corroborate these

Figure 4. Qualitative comparison of 3D GAN inversion methods on the CelebA-HQ [12] dataset. The compared methods include pSp [17], PTI [18], Pose Opt [13], E3DGE [14], IDE-3D [22], and e4e [23]. While all methods produce visually plausible results, noticeable artifacts or identity inconsistencies are observed in varying degrees.

Method	MSE (↓)	LPIPS (↓)	ID Similarity (↑)	MS-SSIM (↓)
e4e [23]	0.05	0.4	0.75	0.38
pSp [17]	0.03	0.17	0.56	0.36
IDE-3D [22]	0.1056	0.2806	0.7194	0.5764
E3DGE [14]	0.097	0.128	0.883	0.220
Pose Opt [13]	**0.0035**	0.0777	0.7013	0.1720
PTI [18]	0.014	0.09	0.9	0.21
SPI [26]	0.0082	**0.0865**	**0.9470**	**0.0991**
Ours	**0.0023**	**0.0689**	**0.975**	**0.0623**

Table 1. Quantitative comparison of reconstruction quality on CelebA-HQ [12] dataset. Lower MSE, LPIPS and MS-SSIM indicate better performance, and higher ID Similarity indicate better identity preservation and perceptual quality. Cells with pink background denote the best results and yellow denote the second best. MS-SSIM is reported as a dissimilarity measure (lower is better), following the convention in PTI [18].

findings. We report most numbers from the original publications and note that differences in data preprocessing or evaluation protocol may affect direct comparability. Our method surpasses all baselines across all metrics, achieving the lowest MSE and LPIPS scores, indicating superior reconstruction accuracy and perceptual similarity. Moreover, our method attains the highest ID Similarity score, reflecting its strong capability in maintaining the subject's identity during inversion. This robustness stems from our method's independence from explicit pose annotations: by performing latent-only optimization within a pose-free generator, our approach preserves fine-grained details and achieves consistent multi-view outputs without relying on external pose estimators or symmetry priors.

4.3. Further Comparisons with DiffPortrait3D and Portrait4D-V2

To contextualize our method beyond the GAN inversion family, we additionally compare with two recent frameworks that address single-image novel-view synthesis through different paradigms: DiffPortrait3D [10], a diffusion-based approach that leverages a pre-trained 2D diffusion prior for zero-shot multi-view portrait synthesis, and Portrait4D-V2 [6], a feed-forward one-shot 4D head synthesizer trained on pseudo multi-view videos.

DiffPortrait3D employs a diffusion-based strategy that leverages a pre-trained 2D diffusion prior for zero-shot multi-view portrait synthesis. Benefiting from the strong generative capability of diffusion models, it produces high-quality images with natural shading and realistic texture even under complex lighting conditions. However, as it lacks explicit geometric supervision, spatial coherence between adjacent views is not strictly enforced—resulting in occasional geometric inconsistencies and inter-frame flicker, especially around occluded or ambiguous regions such as hair or accessories. Portrait4D-V2, in contrast, learns a feed-forward one-shot 4D head synthesizer and

Figure 5. More results of our method. Our method better preserves the subject's identity and facial details, achieving more faithful reconstructions under the same single-view input setting. Here we remove the background for better review of the results.

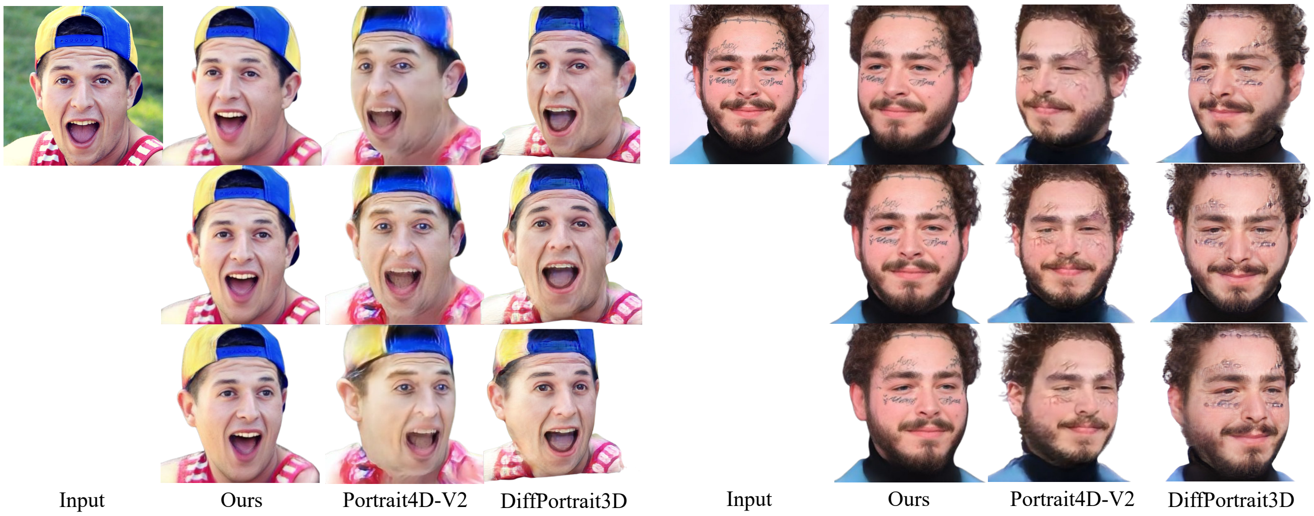

Figure 6. Qualitative comparison with recent portrait generation methods. We evaluate DiffPortrait3D [10] and Portrait4D-V2 [6] on two challenging cases: (*left*) a man wearing a hat, and (*right*) a subject with thin handwritten text on the face. Both baselines produce visually plausible results but exhibit geometric distortions or blurred fine details, while our method maintains clear textual strokes, consistent geometry, and identity stability across novel views.

Method	MSE (↓)	LPIPS (↓)	ID Similarity (↑)	MS-SSIM (↓)
Ours (full)	**0.0023**	**0.0689**	**0.975**	**0.0623**
w/o Pose-appearance consistency D	0.0041	0.0873	0.942	0.089
w/o $\mathcal{L}_{geo}$	0.0037	0.0812	0.955	0.078
w/o $\mathcal{L}_{id}$	0.0028	0.0759	0.928	0.071

Table 2. Ablation study. Removing any component degrades reconstruction and consistency. Lower MSE, LPIPS, and MS-SSIM indicate better quality; higher ID Similarity indicates stronger identity preservation. MS-SSIM is reported as a dissimilarity measure (lower is better), following the convention in PTI [18].

achieves strong temporal consistency as well as controllable pose/expression rendering. It is trained using pseudo multi-view videos constructed from monocular real videos, which reduces reliance on potentially inaccurate 3DMM-based reconstruction. Nevertheless, as a feed-forward model without test-time per-instance optimization, its reconstruction is bounded by the capacity and biases of the learned prior; in challenging cases such as occlusions or fine-grained high-frequency details, it may produce overly smoothed surfaces or locally inconsistent geometry.

To analyze these behaviors in challenging cases, we evaluate both methods on two challenging cases. As shown in Fig. 6, we evaluate these methods on two representative challenging cases that feature fine-grained high-frequency details (handwritten characters) and non-standard facial appearance. Both baselines produce visually plausible results but exhibit limitations in maintaining high-fidelity reconstruction. In contrast, our optimization-based inversion faithfully reconstructs both global structure and fine-grained texture. The handwritten characters remain sharp and spatially consistent under novel-view rotation, while the identity is stably preserved. We note that this comparison is limited to a small number of qualitative examples and that these methods address a broader set of tasks beyond single-image 3D reconstruction; a more comprehensive evaluation is left for future work. By performing instance-specific latent optimization under a pose-free generator design, our method improves structural consistency and identity preservation in these challenging scenarios.

4.4. Ablation Study

We perform ablation experiments to evaluate the contribution of each loss component in our inversion framework. Quantitative results are reported in Table 2, and qualitative comparisons are shown in Fig. 7. All ablation models are evaluated under the same inversion settings unless otherwise specified. We additionally analyze the effect of the pre-trained generator's pose–appearance consistency discriminator on downstream inversion quality.

w/o $\mathcal{L}_{\text{geo}}$. We next evaluate the effect of the geometric regularization term during inversion. Removing $\mathcal{L}_{\text{geo}}$ leads to visible artifacts near facial boundaries and less consistent geometry across viewpoints, as shown in Fig. 7. Correspondingly, MS-SSIM varies noticeably in Table 2, indicating reduced multi-view structural consistency. These results indicate that $\mathcal{L}_{\text{geo}}$ stabilizes the reconstructed geometry and improves structural coherence across viewpoints.

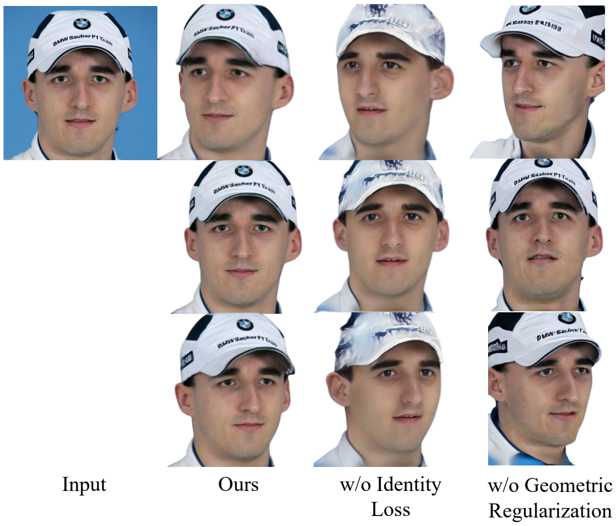

Figure 7. Ablation study on Identity Loss and Geometric Regularization.

w/o $\mathcal{L}_{id}$. Finally, we examine the role of the identity loss. Without $\mathcal{L}_{id}$, reconstructions remain visually plausible but show reduced identity similarity to the input. As reported in Table 2, LPIPS increases while ID similarity decreases when the identity loss is removed. Qualitative results in Fig. 7 further confirm that identity-specific features are less accurately preserved. These results indicate that $\mathcal{L}_{id}$ helps maintain identity fidelity during inversion.

To understand how the quality of the underlying generator affects inversion, we additionally compare against a variant in which the pose–appearance consistency discriminator of [20] is replaced with a standard dual-discriminator architecture that performs only real/fake discrimination without pose conditioning. Without pose-conditioned discrimination, the generator receives no pose-related signal beyond the realism score, resulting in less consistent pose distributions and degraded geometry. Since inversion operates on this pre-trained generator, these quality differences propagate into the inversion stage: as shown in Table 2, reconstruction accuracy and novel-view consistency both degrade when inversion is applied to the weaker generator. This analysis confirms that the quality of pose–appearance disentanglement in the generator is an important factor for downstream inversion performance.

4.5. Limitations

Our method may encounter challenges when handling highly stylized or non-photorealistic images (e.g., Fig. 8), as well as inputs with significant occlusions. This is primarily because the underlying 3D GAN generator is pre-trained on natural human faces, whose training distribution typically lacks complex textures, cartoon-like exaggerations, or severe occlusions. More generally, inversion quality is inherently bounded by the expressiveness of the pre-trained generator. Similar limitations have also been observed in prior GAN inversion approaches that rely on pre-trained generative priors. In addition, as an optimization-based approach, the current inversion process is slower than feed-forward alternatives. Finally, the recovered viewpoint is internally consistent within the generator's canonical frame but does not correspond to a metrically accurate camera pose; applications requiring precise camera calibration would need an additional alignment step.

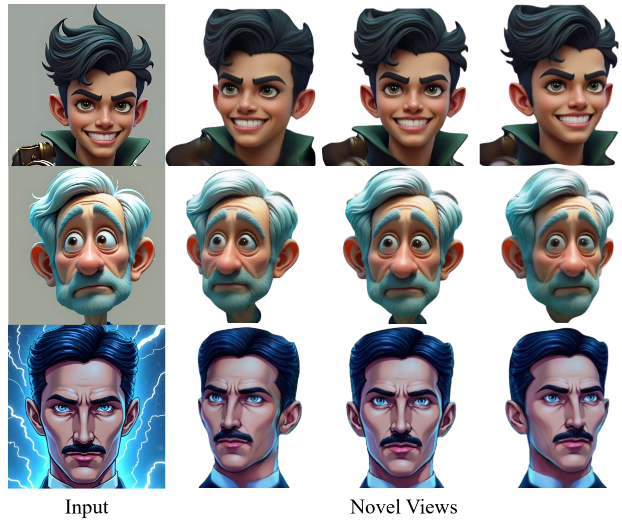

Figure 8. Qualitative results on stylized or exaggerated face images. Our method can handle inputs resembling real human faces (first row), alebit with slight artifacts in the eye region for more cartoonish character(second row).

5. Conclusion

In this paper, we presented an optimization-based 3D GAN inversion framework that improves geometry consistency and identity preservation from single-view inputs. By leveraging a pose-free 3D generator and latent optimization with geometry-aware regularization, our method avoids explicit pose estimation and symmetry assumptions commonly used in prior work. This design enables faithful reconstruction while maintaining stable geometry under diverse viewpoints. Experiments show that our method achieves strong reconstruction quality and multi-view consistency, outperforming existing inversion methods across multiple evaluation metrics. Our framework also demonstrates improved robustness in challenging cases, such as stylized inputs and non-standard facial appearances. In the future, integrating a lightweight encoder for initialization could combine the efficiency of feed-forward methods with the fidelity of optimization-based inversion, enabling real-time applications while preserving 3D consistency.

Acknowledgement

This work was partly supported by EPSRC grant EP/X011364/1 GNOMON

References

[1] R. Abdal, Y. Qin, and P. Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 3

[2] Y. Alaluf, O. Patashnik, and D. Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6711–6720, 2021. 2

[3] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, et al. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5799–5809, 2021. 2

[4] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 1, 2, 3, 4

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 4

[6] Y. Deng, D. Wang, and B. Wang. Portrait4d-v2: Pseudo multi-view data creates better 4d head synthesizer. In *European Conference on Computer Vision*, pages 316–333. Springer, 2024. 6, 8

[7] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 2

[8] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 2

[9] J. Gu, L. Liu, P. Wang, and C. Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 2

[10] Y. Gu, H. Xu, Y. Xie, G. Song, Y. Shi, D. Chang, J. Yang, and L. Luo. Diffportrait3d: Controllable diffusion for zero-shot portrait view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10456–10465, 2024. 6, 8

[11] X. Hu, Q. Huang, Z. Shi, S. Li, C. Gao, L. Sun, and Q. Li. Style transformer for image inversion and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11337–11346, 2022. 2

[12] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5, 6

[13] J. Ko, K. Cho, D. Choi, K. Ryoo, and S. Kim. 3d gan inversion with pose optimization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2967–2976, 2023. 2, 3, 5, 6

[14] Y. Lan, X. Meng, S. Yang, C. C. Loy, and B. Dai. Self-supervised geometry-aware encoder for style-based 3d gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20940–20949, 2023. 1, 3, 5, 6

[15] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y. Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7588–7597, 2019. 2

[16] M. Niemeyer, J. T. Barron, B. Mildenhall, M. Tancik, L. Weber, and A. Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2

[17] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2287–2296, 2021. 2, 5, 6

[18] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1):1–13, 2022. 2, 3, 5, 6, 8

[19] K. Schwarz, Y. Liao, M. Niemeyer, and A. Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 2

[20] Z. Shi, Y. Shen, Y. Xu, S. Peng, Y. Liao, S. Guo, Q. Chen, and D.-Y. Yeung. Learning 3d-aware image synthesis with unknown pose distribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13062–13071, 2023. 2, 3, 4, 9

[21] Z. Shi, Y. Xu, Y. Shen, D. Zhao, Q. Chen, and D.-Y. Yeung. Improving 3d-aware image synthesis with a geometry-aware discriminator. *Advances in Neural Information Processing Systems*, 35:7921–7932, 2022. 2, 3, 4

[22] J. Sun, X. Wang, Y. Shi, L. Wang, J. Wang, and Y. Liu. Ide-3d: Interactive disentangled editing for high-resolution 3d-aware portrait synthesis. *ACM Transactions on Graphics (ToG)*, 41(6):1–10, 2022. 1, 3, 5, 6

[23] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 2, 5, 6

[24] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 5

[25] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang. Gan inversion: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3121–3138, 2022. 1, 2

[26] F. Yin, Y. Zhang, X. Wang, T. Wang, X. Li, Y. Gong, Y. Fan, X. Cun, Y. Shan, C. Oztireli, et al. 3d gan inversion with facial symmetry prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 342–351, 2023. 1, 2, 3, 5, 6

[27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 4, 5