Do Not DeepFake Me: Privacy-Preserving Neural 3D Head Reconstruction Without Sensitive Images

Jiayi Kong[1], Xurui Song[1], Shuo Huai[2], Baixin Xu[1], Jun Luo[2], Ying He[1]*

[1]S-Lab, Nanyang Technological University, Singapore
[2]College of Computing and Data Science, Nanyang Technological University, Singapore
{jiayi006, song0257}@e.ntu.edu.sg, {shuo.huai, baixin001, junluo, yhe}@ntu.edu.sg

Abstract

While 3D head reconstruction is widely used for modeling, existing neural reconstruction approaches rely on high-resolution multi-view images, posing notable privacy issues. Individuals are particularly sensitive to facial features, and facial image leakage can lead to many malicious activities, such as unauthorized tracking and deepfake. In contrast, geometric data is less susceptible to misuse due to its complex processing requirements, and absence of facial texture features. In this paper, we propose a novel two-stage 3D facial reconstruction method aimed at avoiding exposure to sensitive facial information while preserving detailed geometric accuracy. Our approach first uses non-sensitive rear-head images for initial geometry and then refines this geometry using processed privacy-removed gradient images. Extensive experiments show that the resulting geometry is comparable to methods using full images, while the process is resistant to DeepFake applications and facial recognition (FR) systems, thereby proving its effectiveness in privacy protection.

Introduction

In the rapidly evolving digital era, 3D facial reconstruction technology has become an indispensable component in fields such as modeling, virtual reality, and digital entertainment. Recent advancements, particularly the emergence of techniques like Neural Radiance Fields (NeRF) (Mildenhall et al. 2020) and Signed Distance Function (SDF)-based methods (NeuS/VolSDF) (Wang et al. 2021; Yariv et al. 2021), have significantly enhanced the accuracy of reconstruction, heralding a revolutionary progression in this domain. However, these methods often depend on multi-view, high-resolution facial images. While such images provide substantial convenience and precision in reconstruction, they raise significant privacy concerns due to the sensitive nature of the detailed facial information they contain. This reliance on sensitive images inevitably compromises privacy.

Extensive research has consistently demonstrated that individuals exhibit a markedly higher sensitivity to facial features compared to other body parts (Zebrowitz and Montepare 2008). Even minor facial imperfections, such as acne, scars, or uneven skin tone, can exert considerable psychological effects (Hamler et al. 2022; Mekeres et al. 2023). The potential misuse of facial information raises substantial concerns, including risks of social discrimination, psychological distress, and other negative consequences (Abrams et al. 2020). Furthermore, the necessity for protecting facial privacy extends beyond individual sensitivities to include the broader risks associated with the leakage of facial images (Ciftci, Yuksek, and Demir 2023). Current image-based facial recognition (FR) systems have reached a level of sophistication that enables the extraction and analysis of extensive information from facial images (Kortli et al. 2020). In the event of facial image leakage, this data could be exploited for unauthorized digital tracking, surveillance, or other malicious purposes (Hill 2022). Moreover, facial images provide the precise details necessary for creating DeepFake (Chadha et al. 2021), further amplifying privacy risks.

In contrast, geometric data is generally less sensitive compared to facial images. Geometric data alone lacks the detailed facial textures necessary for impersonation and forgery, such as those used in DeepFake technology (Westerlund 2019). This absence of detailed textures significantly diminishes the risk of misuse. Additionally, collecting, processing, and analyzing 3D data is more complex and resource-intensive (Uy et al. 2021), which hinders efficient tracking and recognition. The technology for handling 3D data is less developed and less widespread than that for 2D images (Guo et al. 2023), further reducing the risk of unauthorized monitoring and tracking. Thus, while facial geometry also reveals some information, it is much harder to exploit in DeepFake and FR systems, resulting in fewer malicious effects than the facial images used in current face reconstruction methods. This phenomenon highlights the urgent need to improve current methods and explore high-quality 3D reconstruction techniques that do not depend on high-risk facial images.

Existing privacy protection methods, such as image blurring (Jiang et al. 2023) or facial masking (Sun et al. 2018), while somewhat effective in 2D image processing, face numerous challenges in 3D reconstruction tasks. mm3DFace (Xie et al. 2023) is one of the approaches that address facial privacy in 3D reconstruction by using mmWave signals to extract geometric features, thus avoiding facial images and offering some level of privacy protection.

*Corresponding author

It can track 68 facial landmarks, which is useful for expression analysis. However, it cannot capture geometric details which is necessary for nuanced applications. Currently, there is a lack of a comprehensive end-to-end strategy capable of reconstructing detailed facial geometry without relying on sensitive facial images.

To address this challenge, we propose an innovative 3D neural reconstruction method without dependence on sensitive facial images, representing the first systematic attempt to address privacy protection issues in the field of neural surface reconstruction. Our method employs a two-stage reconstruction process using two types of images: privacy-neutral images (e.q. rear-head images) and privacy-protected images, which are initially facial images processed to remove sensitive information. We first utilize privacy-neutral images to establish the basic geometric structure and then use privacy-protected images to refine and perfect the 3D reconstruction. This approach completely eliminates the reliance on sensitive facial RGB images while still enabling the reconstruction of stable and detailed geometric head models, thus avoiding the risk of sensitive image leakage.

Our research strikes a balance between the requirements of head reconstruction and user privacy protection, offering a reliable and secure head geometry reconstruction solution for various application scenarios. Our main contributions can be summarized as follows:

- We introduce a novel end-to-end neural reconstruction pipeline that effectively protects facial privacy in 3D modeling. Our method eliminates the exposure of sensitive information while maintaining high-quality geometric reconstruction, thereby broadening NeRF's applications in privacy-sensitive scenarios and laying a foundation for privacy-protected advancements.
- We develop a unique two-stage training process that balances reconstruction stability and detail fidelity. This approach first demonstrates that gradient images can effectively contribute to geometric reconstruction, preventing reliance on sensitive facial images.
- Our method significantly improves the security and effectiveness of neural reconstruction in sensitive contexts. Using images that lack detailed facial textures, protects facial privacy and renders them ineffective for common exploitation techniques such as DeepFake creation and facial recognition.

Related Work

Human head models. 3D Morphable Models (3DMM) (Blanz and Vetter 2023) leverage principal component analysis to represent facial. However, this method falls short in capturing details such as wrinkles, the interior of the mouth, and hair, so it may not fully satisfy appearance requirements. i3DMM (Yenamandra et al. 2021) introduces an implicit function to model both the geometry and appearance of the human head with various attributes like shape, expression, and hairstyle. Both approaches rely on 3D data for their modeling. Recent advancements in Neural Radiance Fields (NeRF) (Mildenhall et al. 2020; Barron et al. 2021) have excelled in novel view synthesis due to their compact and powerful representation capability, relying solely on a set of multi-view images. Consequently, numerous works (Gafni et al. 2021; Park et al. 2021) have yielded detailed geometry (Zheng et al. 2022; Xu et al. 2023) and achieved a photo-realistic appearance (Zheng et al. 2023; Kirschstein et al. 2023) in modeling heads.

Neural implicit functions. Sign distance fields (SDF) (Park et al. 2019) and occupancy fields (Mescheder et al. 2019), showcase their representative ability over explicit representations, i.e. mesh, and point cloud. DVR (Niemeyer et al. 2020) and IDR (Yariv et al. 2020) focus on differentiating the surface rendering pipeline based on multi-view images. They incorporate corresponding masks to distinguish objects from the background during the training process. NeuS (Wang et al. 2021) and VolSDF (Yariv et al. 2021) refine the geometry reconstruction of NeRF by introducing a scheme that converts SDF to density, enhancing the surface representation in NeRF. Recent approaches (Rosu and Behnke 2023; Wang, Skorokhodov, and Wonka 2022, 2023) combine volume rendering with multi-scale hashing in Instant-NGP (Müller et al. 2022) and displacement fields to learn detailed surfaces through implicit functions from multi-view images.

Facial privacy protection. We summarize three categories primary approaches. The first involves an algorithm based on adversarial generation networks, aimed at deceiving unauthorized facial recognition by generating fake images highly similar to the original ones (Li, Wang, and Li 2019). However, this method is primarily targeted at public social platforms (Ciftci, Yuksek, and Demir 2023) or specific facial recognition (FR) systems that provide data for learning (Cherepanova et al. 2021). The second category involves the use of cryptographic techniques. Cryptographic techniques, including homomorphic encryption (Huang et al. 2020), secure multiparty computation (Ma et al. 2019), and other encryption primitives (Kou et al. 2021), are used to encrypt original images securely. These approaches introduce higher latency and computational costs. The third group of methods focuses on obfuscation techniques. These methods encompass actions such as implementing blurring (Jiang et al. 2023), introducing noise (Zhang et al. 2021), applying masking (Seneviratne et al. 2022), utilizing filtering (Zhou and Pun 2020), and employing image transformation (Wang, Kelly, and Veldhuis 2021). While practical, these methods irreversibly degrade image quality and may fail in subsequent tasks without a robust end-to-end solution. mm3DFace (Xie et al. 2023) uses mmWave signals for privacy protection, but its goal is not to reconstruct detailed geometry.

Method

Preliminaries

Neural volume rendering. As demonstrated by NeRF (Mildenhall et al. 2020), it characterizes a 3D scene by employing volume density and color fields. Given known camera poses and ray directions, we conduct sampling along the rays, predicting both the color $\mathbf{c}_i$ and

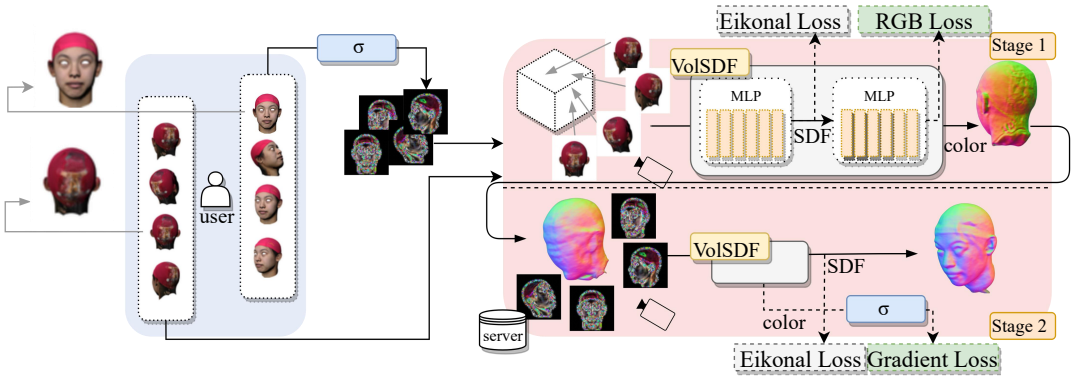


Figure 1: **Algorithmic pipeline**. We show the data flow process where our model reconstructs the human head geometry, without depending on sensitive images. Users can select photos that require privacy protection through a general operator denoted by σ. Following this, privacy-neutral images and privacy-protected images are used for geometric reconstruction. The reconstruction process occurs in two stages: stage 1 uses privacy-neutral images to establish the foundational geometry, while stage 2 employs image gradient information to refine the facial geometry further.

density σ_i at each sample point $\mathbf{x}_i$ using Multi-Layer Perceptron (MLPs). The volume rendering process entails the integrating color radiance across the sampled points along the ray. Consequently, we approximate the rendered color for a specific pixel as:

$$\hat{\mathbf{c}}(\mathbf{o}, \mathbf{d}) = \sum_{i=1}^{N} \omega_i \mathbf{c}_i. \tag{1}$$

In this context, we define the distance of each sample point from the camera center as t_i and introduce δ_i to represent the distance between adjacent sample points: $\delta_i = t_{i+1} - t_i$. Furthermore, we use α_i to quantify the opacity of the i-th ray segment, computed as $\alpha_i = 1 - \exp(-\sigma_i \delta_i)$. The term $T_i = \prod_{j=1}^{i-1}(1 - \alpha_j)$ denotes the accumulated transmittance, indicating the proportion of light that reaches the camera, and ω_i is defined as $\omega_i = T_i \alpha_i$ to determine the in Eq. (1). We train the network using the color loss between the rendered images and input images:

$$\mathcal{L}_{\text{rgb}} = \left\| \hat{\mathbf{c}} - \mathbf{c} \right\|_1. \tag{2}$$

Volume rendering of SDF. One of the most common surface representations is the SDF, which precisely describes an object's geometry. An SDF for a 3D object with a watertight surface is a function $f : \mathbb{R}^3 \to \mathbb{R}$. This function takes any point$\mathbf{x} \in \mathbb{R}^3$ and provides the signed distance between $\mathbf{x}$ and the closest point on the surface. Importantly, the zero-level set of the SDF corresponds to the object's surface S:

$$S = \{\mathbf{x} \in \mathbb{R}^3 \mid f(\mathbf{x}) = 0\}. \tag{3}$$

Recent contributions have been centered around neural volume rendering based on SDF representation (Wang et al. 2021; Yariv et al. 2021). These methods utilize MLPs to implicitly represent a 3D scene by predicting the SDF and color $\mathbf{c}_i$ relative to the viewpoint. This differs from the initial approach presented in (Mildenhall et al. 2020), which focused on predicting color and density. The extraction of the zero-level surface of the SDF in these methods results in a more reasonable geometric representation, showcasing significant effectiveness in reconstructing objects with smooth geometry.

If we treat a function as an SDF, it must adhere to the requirement of differentiability, ensuring that the modulus of the gradient remains constant in accordance with the eikonal equation. Therefore, we incorporate the eikonal loss into the final SDF predictions to ensure that the optimized $f(\mathbf{x}_i)$ conforms to a valid SDF:

$$\mathcal{L}_{\text{eik}} = \frac{1}{N} \sum_{i=1}^{N} (\|\nabla f(\mathbf{x}_i)\|_2 - 1)^2, \tag{4}$$

where N represents the total number of sampled points. Given our use of a network architecture for SDF prediction, computing gradients within a continuous field becomes both feasible and straightforward.

Two-stage Training

In our assumptions, the reconstruction process should neither access nor alter sensitive facial images, ensuring that the algorithm maintains practical utility for geometric reconstruction while fully adhering to privacy requirements. To implement this approach, all images used for reconstruction must be non-sensitive and are categorized into two types: *privacy-neutral* and *privacy-protected*. Neutral pictures, which do not contain facial information, such as those captured from the back of the head, are directly utilized in the training process. Privacy-protected images, initially taken from frontal perspectives, are processed through a specialized operator to ensure they meet our privacy goals before being uploaded by the user and used in the second stage.

In our proposed method, we employ a two-stage training framework for geometric reconstruction in Figure 1. In the first stage, we train a neural radiance field using neutral images, which enables the reconstruction of essential low-frequency information. The success of this stage depends on the quantity of privacy-neutral data users provide. While this

information might have been deemed less valuable for facial geometry learning in the past, it proves to be beneficial in our method. Unless unavoidable circumstances require the use of templates, as discussed in the section **Optimization**, the privacy-neutral images provided by the user can be fully utilized. This enhances the geometric accuracy in the first stage of reconstruction, bringing it closer to the user's geometric features, and ultimately contributing to the overall quality of the head reconstruction in the second stage.

In the second stage, we utilize privacy-protected data uploaded by users for the reconstruction process. Our supervision focuses solely on color variation information inherent in the original images. We intentionally refrain from utilizing full RGB information to mitigate the risk of privacy exposure. While the privacy-protected images appear visually unfriendly and blurry, they contain valuable color variation details crucial for refining geometric intricacies. Building on the foundation established in the first stage, we train for new frontal face viewpoints. We compute gradients of the rendered images, transforming them into multi-view color gradient modulus through a general operator σ. These modulus are then compared to the facial privacy-protected gradient information obtained from the user after passing through the operator. To optimize this stage, we introduce a novel loss equation, the gradient loss $\mathcal{L}_{\text{grad}}$, to guide the learning of geometric information in the second stage:

$$\mathcal{L}_{\text{grad}} = \left\| \|\hat{\mathbf{g}}\|_2 - \|\mathbf{g}\|_2 \right\|_1, \tag{5}$$

where $\mathbf{g}$ and $\hat{\mathbf{g}}$ are color gradients of the original image and the predicted gradient information, respectively. We calculate the gradient of the color in both $x-$ and $y-$ directions and obtain their modulus as follows:

$$\|\mathbf{g}\|_2 = \left\| \frac{\partial \mathbf{c}}{\partial x} \right\|_2 + \left\| \frac{\partial \mathbf{c}}{\partial y} \right\|_2. \tag{6}$$

In practical implementation, this operator can take the form of an edge extraction operator, such as the Sobel operator. Its design aims for versatility, enabling users to choose from a range of operators σ depending on their specific privacy requirements and to adjust operator parameters to ensure the irreproducibility of the reconstruction. This flexibility empowers users to acquire data with different levels of protection. During this stage, the absence of RGB supervision may lead to inaccuracies in MLP color predictions. Nevertheless, our method effectively recovers facial geometric details even when radiance information appears unreliable.

The privacy-preserving approach focuses on three key aspects: irreversibility by retaining only gradient magnitudes, color multiplicity where different images can map to the same gradient, and perceptual indistinction to keep the processed data visually indistinguishable from the original. Detailed discussions are in the supplementary material.

Optimization

Template. Templates offer an effective solution when users cannot capture photos without privacy-sensitive information, such as handheld devices or other constraints. In such cases, user-uploaded images need processing for privacy protection. To address this challenge, we propose using a neutral head template. This enables users to recover head geometry even when privacy-neutral images are scarce, facilitating effective geometric reconstruction. While template usage is not mandatory, it enhances our model's compatibility with various data inputs, making it more versatile for a smoother user experience.

Regularization. In our network structure, based on the neural radiance fields pipeline, we make specific optimizations to enhance geometric representation. Our primary goal is not the final rendering appearance but the recovery of fine geometric details for downstream tasks. To achieve this, we apply regularization constraints to the color rendering network. These constraints encourage the network to prioritize learning geometric intricacies. In previous work (Rosu and Behnke 2023), a color regularization constraint was proposed which introduced trainable bounds for each layer, effectively constraining the expression of MLP layers using knowledge from Lipschitz continuous networks. In the specific network structure, each MLP layer is reformulated as $y = \sigma(\widehat{W}_i \mathbf{x} + b_i)$, and $\widehat{W}_i = a(W_i, \text{softplus}(m_i))$, where a normalizes the weight matrix. During the privacy-protected stage, we apply color regularization by constraining the product of Lipschitz constants, m_i, for each layer:

$$\text{softplus}(m_i) = \ln\left(1 + e^{m_i}\right). \tag{7}$$

In the training process for stage one, we do not use this regularization term. In the second stage of training, we introduce constraints on the rendering network as an additional loss:

$$\mathcal{L}_{\text{lip}} = \prod_{i=1}^{l} \text{softplus}(m_i). \tag{8}$$

Putting it all together, our training loss is as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{rgb}} + \lambda_2 \mathcal{L}_{\text{eik}} + \lambda_3 \mathcal{L}_{\text{lip}} + \lambda_4 \mathcal{L}_{\text{grad}}. \tag{9}$$

In Stage 1, we set $\lambda_3 = \lambda_4 = 0$ to disable gradients, and set $\lambda_1 = 0$ in Stage 2 to activate it.

Experiments

Setup

Datasets. In our experiments, we utilize two representative datasets: FaceScape (Yang et al. 2020) and High-Fidelity 3D Head (H3DS) (Ramon et al. 2021). Each dataset includes 30 to 36 RGB images per identity at 64×64 pixels, with lower resolution and increased blurring chosen to enhance privacy protection. FaceScape offers textured 3D face data for various subjects and expressions, while H3DS, captured in real-world scenarios, provides headshot data across different countries, ethnic backgrounds, and lighting conditions. We randomly sampled 10 identities from each dataset and treated all images with visible facial features as sensitive images. The selection of images for training is adaptable to user privacy preferences, as detailed in the supplementary material. In our experiments with the FaceScape dataset, we utilize 10 images that are inherently privacy-neutral and process 20 images as privacy-protected. Similarly, for the H3DS dataset, 16 images are inherently privacy-neutral, and 20 images are processed as privacy-protected. While our primary

analysis focuses on these two datasets, we also test additional facial datasets to broaden our validation in the supplementary. These experiments demonstrate the adaptability and robustness of our method in various conditions.

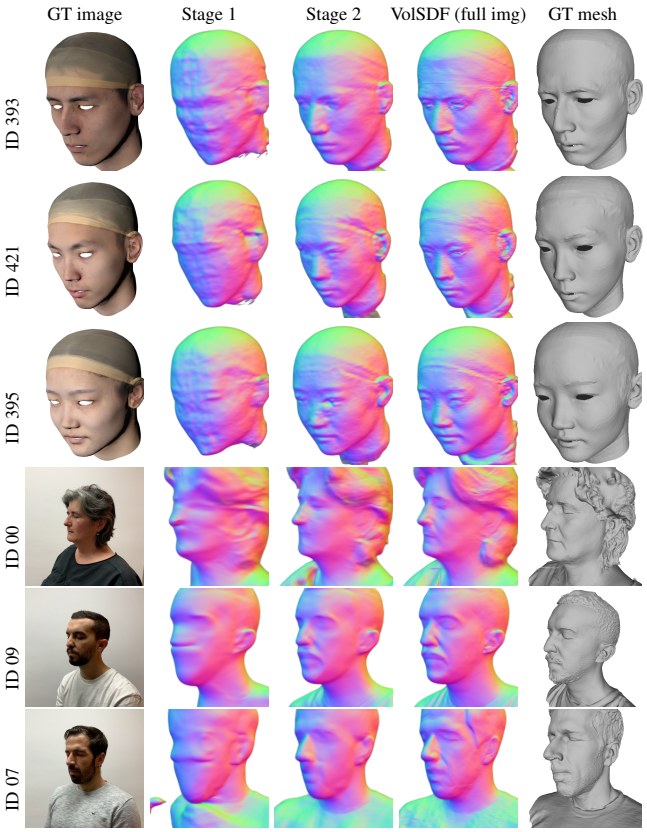


Figure 2: A two-stage strategy: stage 1 focuses on learning the basic geometry of the head, and stage 2 only uses gradient images to optimize facial geometric details. Even compared with the full RGB image given 30 viewing directions, our reconstruction still achieves comparable results.

Baseline. Our approach is based on VolSDF (Yariv et al. 2021) due to its stability in constructing detailed facial geometry. The flexibility of our framework allows for the incorporation of alternative rendering and training methods such as 2DGS (Huang et al. 2024) in Figure 10. We evaluate the geometric accuracy of our work under identical camera poses and view directions. VolSDF is based on all full-face images, while our method doesn't rely on sensitive facial images. Given that our method is trained on gradient images, a large amount of information is lost compared to full RGB images at all viewing angles, and our results remain comparable to VolSDF, as detailed in the Evaluation.

Evaluation. We evaluate the quality of our algorithm's geometry recovery through both qualitative and quantitative comparisons with the VolSDF algorithm. We present results obtained after the first stage of privacy-neutral training and the second stage of privacy-protected reconstruction, comparing them with VolSDF's full-image supervised

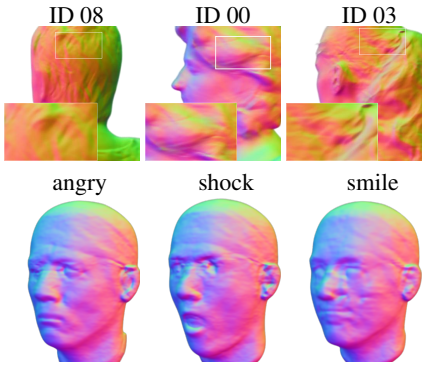
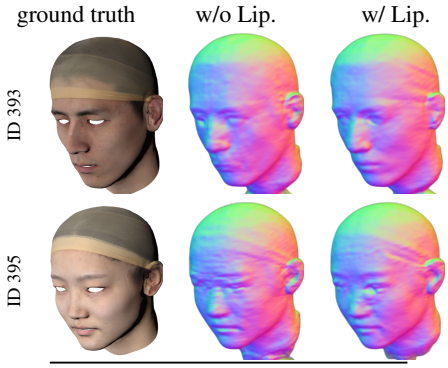


Figure 3: In terms of reconstruction quality, even with 64-resolution images, we are still able to capture many geometric details, including hair and expressions.



CD (10^{-3})	ID 393	ID 395
w/ Lip.	0.6998	0.7664
w/o Lip.	0.7753	0.8347

Figure 4: Lipschitz regularization (Lip.) makes geometry learning more reasonable in Stage 2, such as the mouth (Row 1) and the smoother face (Row 2).

reconstruction outcomes in Figure 2. To evaluate the overall reconstruction accuracy, we employ the Chamfer distance (CD) metric, computed for both our method and VolSDF. These metrics are detailed in Table 1. Furthermore, our reconstruction excels in capturing intricate details such as hair or interesting facial features such as different expressions in Figure 3. This highlights the versatility of our method in handling intricate aspects beyond facial contours.

Ablation Studies

Lipschitz regularization. We conduct an ablation study of the Lipschitz regularization as depicted in Figure 4. Since the face represents relatively smooth geometry, Lipschitz regularization helps ensure more accurate geometric learning and plays a crucial role in recovering facial details, ultimately resulting in the lowest CD.

Sparse views and geometric priors. Considering varying privacy definitions, we compare scenarios with fewer color images in Stage 1 (Figure 5). Stricter privacy requirements may impact Stage 1 accuracy, but the final results remain

Method↓	393	421	395	346	340	375	411	393_4	393_3	393_2	**Mean**
Ours stage 1	2.7692	1.7467	2.8889	2.3320	1.7957	2.6392	2.2117	2.1790	2.9775	2.0121	2.3552
Ours stage 2	0.6998	0.7176	0.7645	0.7137	0.8382	0.8289	0.8294	0.8556	0.7798	0.7965	0.7824
VolSDF (Yariv et al. 2021)(full img)	0.4509	0.5163	0.5399	0.4547	0.4988	0.5266	0.5059	0.4366	0.5394	0.4788	0.4948

(a) CD (10^{-3}) results on 10 identities in the FaceScape dataset (Yang et al. 2020).

Method↓	00	01	02	03	04	05	06	07	08	09	**Mean**
Ours stage 1	3.8912	3.9796	3.508	4.1967	3.8524	3.8205	5.9922	6.8757	4.4531	5.0861	4.7066
Ours stage 2	3.0397	3.1610	2.8901	3.4117	2.9827	3.6830	3.3911	3.0951	3.1113	2.5912	3.0477
VolSDF (Yariv et al. 2021) (full imgs)	2.7931	2.2665	2.3108	2.5131	2.607	2.6912	3.1767	2.7124	2.9878	2.3938	2.6351

(b) CD (mm) results on 10 identities in the H3DS dataset (Ramon et al. 2021).

Table 1: We utilize CD to evaluate the quality of the reconstructed mesh, where lower values indicate better performance. To ensure a fair comparison, we employ the same technique to trim and process the mesh. In comparison to the first stage, our geometric accuracy has improved, and the CD value has been reduced to a level comparable to that of the full RGB image input.

acceptable. In extreme cases where all provided images contain sensitive features, a one-stage geometric simulation using identity-free templates trained on multiple heads can be used. Figure 6 shows that while these templates can't fully replace original reconstructions due to missing low-frequency data, they are effective in specific situations.

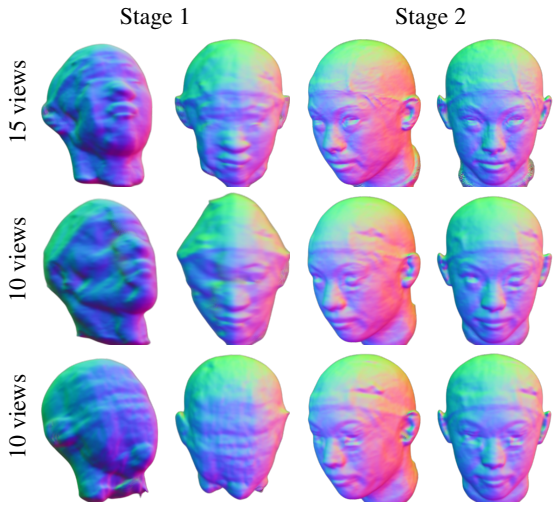


Figure 5: Varying the number of RGB images and views in the first stage may affect Stage 1 accuracy but has minimal impact on final reconstruction quality. Row 1 shows a one-stage process with 15 views, while Rows 2 and 3, both with 10 views, display different input views.

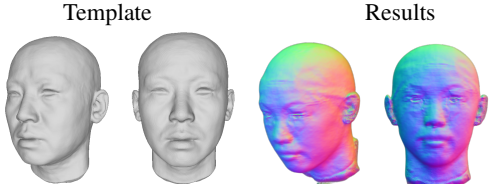


Figure 6: Template. Two on the left showcase the neutral template we employ, while two on the right are our reconstruction results using the template.

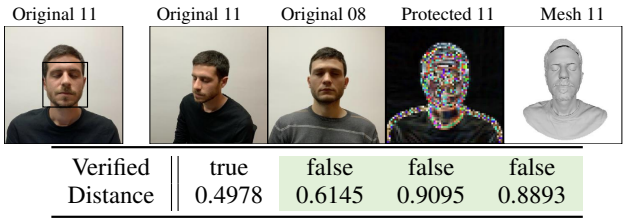
Our experiments show that the first stage is crucial for extracting essential geometric priors, benefiting the second stage significantly. Without this initial stage, reconstructions from gradient examples lack smoothness and plausible geometry, yield unsatisfactory results. This occurs because inadequate low-frequency information during the initial training disrupts the learning process when high-frequency data is introduced.

Method	Original	Protected	Rendered Mesh	Our Rendering	VolSDF Rendering
VGG-Face	93.75	6.51	3.03	43.74	84.62
Facenet	96.88	9.68	18.18	31.23	81.52
Facenet512	96.87	3.23	3.20	18.77	79.27
ArcFace	90.63	6.66	6.46	31.25	79.23
Dlib	93.75	3.17	1.77	37.50	71.57
SFace	84.38	6.45	9.14	43.71	73.35
GhostFaceNet	85.42	6.62	1.04	31.44	73.92
Average	91.67	6.05	6.12	33.95	77.64

Table 2: Comparison of recognition accuracy (%) across 7 FR systems shows that our protected images, directly rendered mesh images, and our rendered results all significantly reduce recognition accuracy, thus avoiding FR tracking.

Security Discussion

Avoiding 2D FR tracking. Although 3D meshes contain identifiable information, they are more robust against malicious exploitation compared to 2D facial images. This robustness stems from the challenges in collecting large-scale 3D datasets, which has limited the development of deep learning-based 3D face recognition (Guo et al. 2023). In contrast, most widely deployed FR systems rely on 2D images due to their speed and convenience. The powerful capabilities of 2D FR systems, if misused for tracking and surveillance, can lead to severe privacy concerns. To demonstrate the effectiveness of our privacy-preserving method, we evaluated its resistance against seven state-of-the-art (SOTA) 2D face recognition systems (Parkhi, Vedaldi, and Zisserman 2015; Schroff, Kalenichenko, and Philbin 2015; Deng et al. 2019; King 2009; Zhong et al. 2021; Alansari et al. 2023), which typically represent faces as vectors, using metrics like cosine similarity (Figure 7) to measure the similarity between images of the same person. The tests results in Table 2 included original sensitive images (Original), pro-



Verified	true	false	false	false
Distance	0.4978	0.6145	0.9095	0.8893

Figure 7: VGG verification of the four images (right) against the original (left), showing cosine distances. Greater distances indicate lower similarity.

tected images (Protected), rendering results from both our method (Our Rendering) and VolSDF (VolSDF Rendering), and images directly rendered from 3D geometric meshes (Rendered Mesh). which confirms that our approach significantly reduces FR accuracy, demonstrating its effectiveness in preventing tracking and lowering security risks.

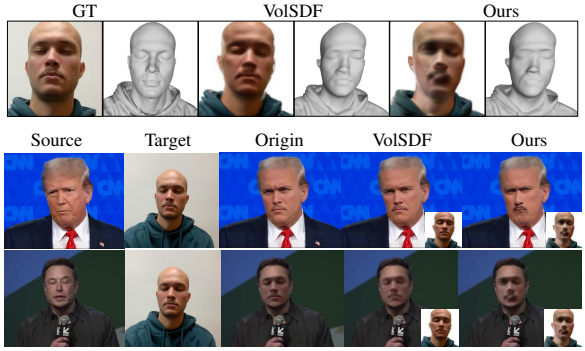


Figure 8: We reconstruct geometric details without rendering realistic sensitive images (row 1). Our rendered images after DeepFake processing retain significantly different facial details from the target person (row 2,3), showing our method prevents convincing impersonation.

Counteracting DeepFake. Current DeepFake technology can convincingly replace facial details through other people's images and videos, resulting in highly realistic and potentially harmful fabrications. In NeRF-based reconstructions, DeepFake can participate in fabrications in two ways: one requires users to provide multi-view facial images, which our method avoids, and the other applies DeepFake to rendered images. Figure 8 compares results from DeepFake models with (Row 2) and without (Row 3) pre-trained generative models, which inputs include original multi-view RGB images, images rendered with VolSDF, and images rendered using our method. Results show that DeepFake images generated with our method fail to convincingly replicate the target individual, making DeepFake ineffective due to the loss of sensitive facial textures.

Preventing misuse. We investigated the role of facial texture in visual recognition by transferring the radiance and geometry from different identities. Figure 9 demonstrated that facial mesh transferred with someone else's facial texture radiance appeared significantly more similar to that person, reinforcing the idea that achieving accurate forgery of facial identities requires the acquisition of highly specific, original facial textures. Our findings highlight that without access to the exact texture of the original face, the potential for successful face forgery remains limited.

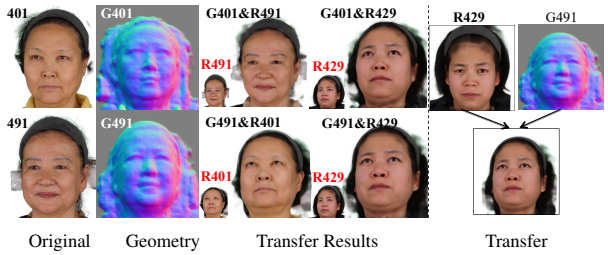


Figure 9: We use $G*$ for geometry and $R*$ for radiance of identity $*$. Each row shows the same geometry with varying radiance, demonstrating that radiance significantly impacts perceived identity.

Future application. Our method prevents reliance on sensitive images by achieving high-fidelity geometric reconstruction using only sensitive-image gradients as inputs. This approach extends beyond facial reconstruction to scenes requiring minimal input, such as meeting rooms or protected elements like license plates (Figure 10), while maintaining comparable quality. It is applicable to other neural reconstruction pipelines (Huang et al. 2024). Future work will focus on developing flexible adjustments to balance privacy and quality for diverse applications and user needs.

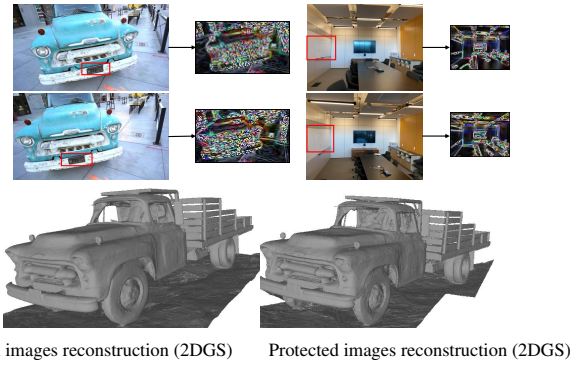


Figure 10: For sensitive images (license plates, conference room details), users can provide only protected inputs instead of the actual images, and reconstruct details.

Conclusion

In this paper, we highlight the often-overlooked aspect of privacy in neural facial reconstruction. We present a method that reconstructs detailed head geometry without relying on sensitive input. By processing sensitive facial data and focusing on essential geometric information in a two-stage process, our approach balances privacy protection with reconstruction quality. Our method represents a significant advancement, integrating privacy and neural facial reconstruction and paving the way for new explorations in this field.

Acknowledgments

This work was supported in part by the Ministry of Education, Singapore, under its Academic Research Fund Grants (MOE-T2EP20220-0005 & RT19/22) and the RIE2020 Industry Alignment Fund–Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

Abrams, J. A.; Belgrave, F. Z.; Williams, C. D.; and Maxwell, M. L. 2020. African American adolescent girls' beliefs about skin tone and colorism. *Journal of Black Psychology*, 46(2-3): 169–194.

Alansari, M.; Hay, O. A.; Javed, S.; Shoufan, A.; Zweiri, Y.; and Werghi, N. 2023. Ghostfacenets: Lightweight face recognition model from cheap operations. *IEEE Access*, 11: 35429–35446.

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. *ICCV*.

Blanz, V.; and Vetter, T. 2023. A morphable model for the synthesis of 3D faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 157–164.

Chadha, A.; Kumar, V.; Kashyap, S.; and Gupta, M. 2021. Deepfake: an overview. In *Proceedings of second international conference on computing, communications, and cyber-security: IC4S 2020*, 557–566. Springer.

Cherepanova, V.; Goldblum, M.; Foley, H.; Duan, S.; Dickerson, J. P.; Taylor, G.; and Goldstein, T. 2021. LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ciftci, U. A.; Yuksek, G.; and Demir, I. 2023. My face my choice: Privacy enhancing deepfakes for social media anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1369–1379.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4690–4699.

Gafni, G.; Thies, J.; Zollhofer, M.; and Nießner, M. 2021. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8649–8658.

Guo, Y.; Wang, H.; Wang, L.; Lei, Y.; Liu, L.; and Bennamoun, M. 2023. 3d face recognition: Two decades of progress and prospects. *ACM Computing Surveys*, 56(3): 1–39.

Hamler, T. C.; Nguyen, A. W.; Keith, V.; Qin, W.; and Wang, F. 2022. How skin tone influences relationships between discrimination, psychological distress, and self-rated mental health among older African Americans. *The Journals of Gerontology: Series B*, 77(11): 2026–2037.

Hill, K. 2022. The secretive company that might end privacy as we know it. In *Ethics of Data and Analytics*, 170–177. Auerbach Publications.

Huang, B.; Yu, Z.; Chen, A.; Geiger, A.; and Gao, S. 2024. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Huang, Y.; Song, Z.; Li, K.; and Arora, S. 2020. Instahide: Instance-hiding schemes for private distributed learning. In *International conference on machine learning*, 4507–4518. PMLR.

Jiang, B.; Bai, B.; Lin, H.; Wang, Y.; Guo, Y.; and Fang, L. 2023. DartBlur: Privacy Preservation With Detection Artifact Suppression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16479–16488.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10: 1755–1758.

Kirschstein, T.; Qian, S.; Giebenhain, S.; Walter, T.; and Nießner, M. 2023. NeRSemble: Multi-view Radiance Field Reconstruction of Human Heads. *ACM Trans. Graph.*, 42(4): 161:1–161:14.

Kortli, Y.; Jridi, M.; Al Falou, A.; and Atri, M. 2020. Face recognition systems: A survey. *Sensors*, 20(2): 342.

Kou, X.; Zhang, Z.; Zhang, Y.; and Li, L. 2021. Efficient and privacy-preserving distributed face recognition scheme via facenet. In *Proceedings of the ACM Turing Award Celebration Conference-China*, 110–115.

Li, Y.; Wang, Y.; and Li, D. 2019. Privacy-preserving lightweight face recognition. *Neurocomputing*, 363: 212–222.

Ma, Z.; Liu, Y.; Liu, X.; Ma, J.; and Ren, K. 2019. Lightweight privacy-preserving ensemble classification for face recognition. *IEEE Internet of Things Journal*, 6(3): 5778–5790.

Mekeres, G. M.; Buhaș, C. L.; Csep, A. N.; Beiușanu, C.; Andreescu, G.; Marian, P.; Cheregi, C. D.; Fodor, R.; and Manole, F. 2023. The importance of psychometric and physical scales for the evaluation of the consequences of scars—A literature review. *Clinics and Practice*, 13(2): 372–383.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4460–4470.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4): 1–15.

Niemeyer, M.; Mescheder, L.; Oechsle, M.; and Geiger, A. 2020. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3504–3515.

Park, J. J.; Florence, P.; Straub, J.; Newcombe, R.; and Lovegrove, S. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 165–174.

Park, K.; Sinha, U.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Seitz, S. M.; and Martin-Brualla, R. 2021. Nerfies: Deformable Neural Radiance Fields.

Parkhi, O.; Vedaldi, A.; and Zisserman, A. 2015. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association.

Ramon, E.; Triginer, G.; Escur, J.; Pumarola, A.; Garcia, J.; Giro-i Nieto, X.; and Moreno-Noguer, F. 2021. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5620–5629.

Rosu, R. A.; and Behnke, S. 2023. Permutosdf: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8466–8475.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Seneviratne, S.; Kasthuriarachchi, N.; Rasnayaka, S.; Hettiachchi, D.; and Shariffdeen, R. 2022. Does a face mask protect my privacy?: Deep learning to predict protected attributes from masked face images. In *Australasian Joint Conference on Artificial Intelligence*, 91–102. Springer.

Sun, Q.; Ma, L.; Oh, S. J.; Van Gool, L.; Schiele, B.; and Fritz, M. 2018. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5050–5059.

Uy, M. A.; Kim, V. G.; Sung, M.; Aigerman, N.; Chaudhuri, S.; and Guibas, L. J. 2021. Joint learning of 3d shape retrieval and deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11713–11722.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*.

Wang, S.; Kelly, U. M.; and Veldhuis, R. N. 2021. Gender obfuscation through face morphing. In *2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*, 1–6. IEEE.

Wang, Y.; Skorokhodov, I.; and Wonka, P. 2022. Hf-neus: Improved surface reconstruction using high-frequency details. *Advances in Neural Information Processing Systems*, 35: 1966–1978.

Wang, Y.; Skorokhodov, I.; and Wonka, P. 2023. PET-NeuS: Positional Encoding Tri-Planes for Neural Surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12598–12607.

Westerlund, M. 2019. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11).

Xie, J.; Kong, H.; Yu, J.; Chen, Y.; Kong, L.; Zhu, Y.; and Tang, F. 2023. mm3DFace: Nonintrusive 3D Facial Reconstruction Leveraging mmWave Signals. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, 462–474.

Xu, B.; Zhang, J.; Lin, K.-Y.; Qian, C.; and He, Y. 2023. Deformable Model-Driven Neural Rendering for High-Fidelity 3D Reconstruction of Human Heads Under Low-View Settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 17924–17934.

Yang, H.; Zhu, H.; Wang, Y.; Huang, M.; Shen, Q.; Yang, R.; and Cao, X. 2020. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 601–610.

Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34: 4805–4815.

Yariv, L.; Kasten, Y.; Moran, D.; Galun, M.; Atzmon, M.; Ronen, B.; and Lipman, Y. 2020. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33: 2492–2502.

Yenamandra, T.; Tewari, A.; Bernard, F.; Seidel, H.-P.; Elgharib, M.; Cremers, D.; and Theobalt, C. 2021. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12803–12813.

Zebrowitz, L. A.; and Montepare, J. M. 2008. Social psychological face perception: Why appearance matters. *Social and personality psychology compass*, 2(3): 1497–1517.

Zhang, X.; Ding, J.; Wu, M.; Wong, S. T.; Van Nguyen, H.; and Pan, M. 2021. Adaptive privacy preserving deep learning algorithms for medical data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1169–1178.

Zheng, M.; Haiyu, Z.; Yang, H.; and Huang, D. 2023. NeuFace: Realistic 3D Neural Face Rendering from Multi-view Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zheng, Y.; Abrevaya, V. F.; Bühler, M. C.; Chen, X.; Black, M. J.; and Hilliges, O. 2022. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13545–13555.

Zhong, Y.; Deng, W.; Hu, J.; Zhao, D.; Li, X.; and Wen, D. 2021. Sface: Sigmoid-constrained hypersphere loss for robust face recognition. *IEEE Transactions on Image Processing*, 30: 2587–2598.

Zhou, J.; and Pun, C.-M. 2020. Personal privacy protection via irrelevant faces tracking and pixelation in video live streaming. *IEEE Transactions on Information Forensics and Security*, 16: 1088–1103.