1D-DiffNPHR: 1D Diffusion Neural Parametric Head Reconstruction Using a Single Image

Pragati Jaiswal[1,2], Tewodros Amberbir Habtegebrial[1,2] and Didier Stricker[1,2]
[1] *RPTU, Technische Universität Kaiserslautern, Germany*
[2] *DFKI, German Research Center for Artificial Intelligence, Germany*

Keywords: Single-View Reconstruction, 3D Face Reconstruction, 1D-Diffusion.

Abstract: In the field of 3D reconstruction, recent developments, especially in face reconstruction, have shown considerable promise. Despite these achievements, many of these techniques depend heavily on a large number of input views and are inefficient limiting their practicality. This paper proposes a solution to these challenges by focusing on single-view, full 3D head reconstruction. Our approach leverages a 1D diffusion model in combination with RGB image features and a neural parametric latent representation. Specifically, we train a system to learn latent codes conditioned on features extracted from a single input image. The model directly processes the input image at inference to generate latent codes, which are then decoded into a 3D mesh. Our method achieves high-fidelity reconstructions that outperform state-of-the-art approaches such as 3D Morphable Models, Neural Parametric Head Models, and existing methods for head reconstruction.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

3D head reconstruction has become a pivotal area in computer vision due to its applications in virtual/augmented reality, medical imaging, and surgical planning. Traditional multi-view reconstruction techniques (Furukawa and Ponce, 2009; Schönberger et al., 2016; Campbell et al., 2008) rely on feature matching across multiple images, requiring extensive image overlap and computational resources. While effective, their high computational cost limits applicability, particularly in single-view scenarios.

To overcome these limitations, single-view approaches such as 3D Morphable Models (3DMMs) (Blanz and Vetter, 1999) use PCA to represent facial geometry in a low-dimensional space, enabling reconstructions from a single image. Advanced models like DECA (Feng et al., 2021), MICA (Zielonka et al., 2022), HI-Face (Chai et al., 2023), and HRN (Lei et al., 2023) have improved reconstruction quality but struggle with full-head details, including complex features like hair and nuanced expressions, due to the inherent limitations of PCA. To address the limitations of 3DMMs, Neural Parametric Head Model (NPHM) (Giebenhain et al., 2023) offers significant advantages in 3D head reconstruction by explicitly learning deformable head shapes and facial expressions but it depends on 3D point cloud data. In contrast, the Monocular Neural Parametric Head Model (MonoNPHM) (Li et al., 2023) generates high-fidelity 3D heads from monocular video, capturing dynamic expressions but its reliance on optimization in 2D space decreases the accuracy. The 3D is estimated using the loss in 2D rendered space, which can result in inconsistencies with the geometry and appearance, particularly when compared to 3D ground truth.

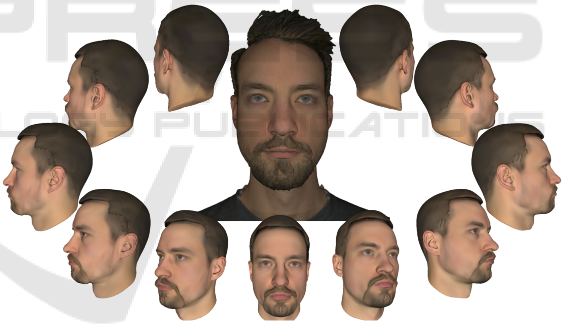


Figure 1: We introduce **1D-DiffNPHR**, a high-fidelity 3D head reconstruction method using a single image. The figure shows the 360° view of the reconstructed 3D geometry given an image.

Diffusion-based models, such as Morphable Diffusion (Wang et al., 2024) and Diffusion Rig (Zheng Ding and Zhang, 2023), focus on achieving efficient novel view synthesis by generating consistent 3D-aware images. While methods like Rodin

Jaiswal, P., Habtegebrial, T. A. and Stricker, D.
1D-DiffNPHR: 1D Diffusion Neural Parametric Head Reconstruction Using a Single Image.
DOI: 10.5220/0013310300003905
In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2025)*, pages 393-400
ISBN: 978-989-758-730-6; ISSN: 2184-4313
Copyright © 2025 by Paper published under CC license (CC BY-NC-ND 4.0)

(Zhao et al., 2022) and Rodin HD (Zhao et al., 2023) employ triplane-based approaches to produce high-quality 3D-aware synthesis while aiming to optimize efficiency compared to traditional volumetric rendering. However, the potential of using a more efficient approach is still open. This limitation motivates our approach, which introduces a diffusion-based model using a 1D diffusion architecture to streamline the reconstruction process.

We introduce **1D-DiffNPHR**, a novel method for high-fidelity 3D head reconstruction from a single image. Our contributions are as follows:

- **1D Diffusion for Full-Head Reconstruction:** Our 1D diffusion model achieves state-of-the-art (SOTA) accuracy and inference time in reconstructing detailed full 3D head geometry, including facial expressions and its transfer from a single RGB image.
- **Enhanced Feature Integration:** By directly integrating RGB image features into the diffusion process and calculating losses in a compact 1D latent space, we eliminate the need for complex 2D or 3D loss computations, optimizing the training.

2 RELATED WORK

2.1 3D Morphable Models

3DMMs have been pivotal for facial reconstruction, using linear subspaces derived from 3D scans to define shape and texture (Cao et al., 2013; Booth et al., 2017; Booth et al., 2016). Subsequent advancements leveraged larger datasets and enhanced statistical models for controllable expressions and finer details (Ploumpis et al., 2019; Tran et al., 2019; Tran and Liu, 2018; Li et al., 2017). Despite progress, 3DMMs struggle with representing complex topologies like full heads and diverse hairstyles, requiring more flexible approaches. Neural Head (Grassal et al., 2022) addresses this by combining 3DMM priors with deep neural networks for high-frequency detail. However, PCA's inherent smoothing limits 3DMMs' ability to capture realistic and diverse facial structures.

2.2 Neural Parametric Head Model

NPHM and Deferred Diffusion (Kirschstein et al., 2023) demonstrate significant potential in 3D head reconstruction. NPHM excels in high-quality 3D reconstruction by learning deformable head shapes and expressions, but its reliance on 3D point cloud data can hinder practicality. Deferred Diffusion integrates multi-view video data with diffusion techniques for view consistency but requires substantial effort in data capture and preprocessing. Both methods face challenges in real-world scenarios due to high data requirements and preprocessing demands.

2.3 Diffusion-Based Models

Diffusion models have shown promise in generating 3D. Techniques like DreamFusion (Poole et al., 2022) and Diffrf (Müller et al., 2023) generate 3D objects as radiance fields, but often lack fine detail and realistic texture. Control3Diff (Gu et al., 2023) uses 2D diffusion to sample triplanes for multiview rendering but struggles with resolution, occlusion, and view inconsistencies.

In contrast, 1D diffusion models excel in spatio-temporal applications like time-series forecasting and trajectory generation, as demonstrated by RecFusion (Bénédict et al., 2023) and DiffTraj (Zhu et al., 2023b). While effective in 1D contexts, these models have not been explored for 3D reconstruction. Our work extends the concept of 1D diffusion by applying it to full 3D head reconstruction, leveraging the simplicity and efficiency of 1D diffusion to overcome the complexity of higher-dimensional diffusion models in 3D facial and head modelling.

3 METHOD

We introduce a 3D head reconstruction pipeline as shown in Fig. 2, which takes an RGB image to produces a high-fidelity 3D head mesh that can be used directly in standard graphics pipelines. Leveraging pre-trained 2D facial models for visual details and latent representations for geometry, the approach generates a morphable 3D head suitable for animations and deformations.

3.1 Face Image Embeddings

Given an input image I with dimensions $512 \times 512 \times 3$, we utilize FaRL (Face Representation Learning) (Zheng et al., 2021) we generate a conditional embedding $\mathbf{z}_{\text{cond}} \in \mathbb{R}^d$, where d is the dimension of the embedding vector (in our case, $d = 512$). This embedding encodes critical facial attributes, including identity and expression-specific features, essential for guiding the diffusion process of 3D head reconstruction.

$$\mathbf{z}_{\text{cond}} = M(I) \tag{1}$$

where $M(I)$ is the function for obtaining the conditional embedding from the input image.

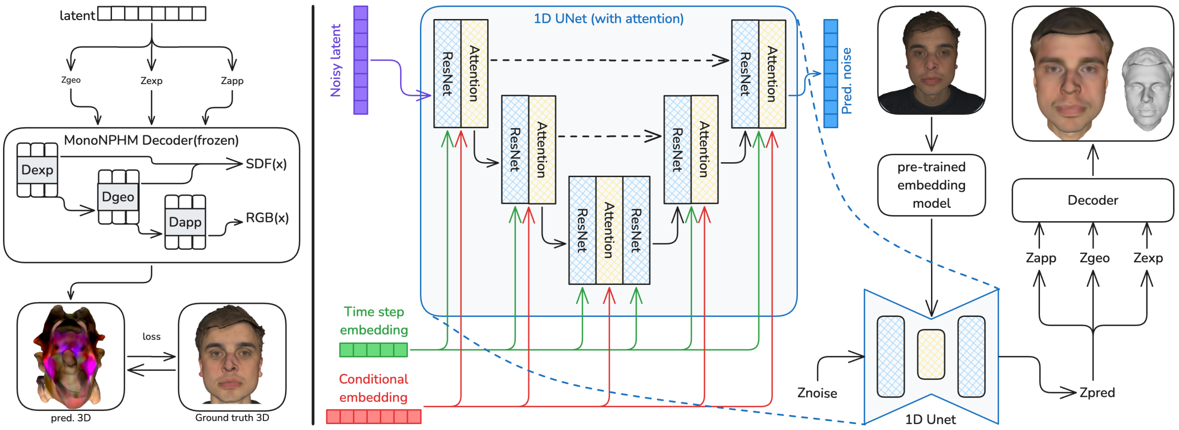


Figure 2: **1D-DiffNPHR:** The pipeline for 3D head reconstruction using a 1D diffusion model. On the left is the illustration of the process for deriving ground truth latent codes (z_{geo}, z_{app}, and z_{exp}). They are randomly initialized and iteratively optimized using a frozen decoder. On the right, we have the **training phase** employing a 1D U-Net with attention to denoise latent representations, guided by time-step and conditional embeddings from a pre-trained embedding model. During **inference phase**, the diffusion model generates latents from the input image, which are decoded to reconstruct a detailed 3D head.

3.2 Conditional 1D Latent Diffusion

Our model employs a 1D U-Net with attention layers to generate latent $\mathbf{z}_{\text{true}} \in \mathbb{R}^{4456\times1}$ (see Section 4.1.1) guided by a conditional embedding $\mathbf{z}_{\text{cond}} \in \mathbb{R}^{512\times1}$. At each timestep t, the U-Net refines the latent using:

Latent representation $\mathbf{z}_t$: Refined progressively to reduce noise.

Conditional embedding $\mathbf{z}_{\text{cond}}$: Derived from the input image I to ensure retention of unique facial features of each individual.

Time-step embedding $\gamma(t)$: Encodes timestep for noise management.

The U-Net processes $\mathbf{z}_t$ via attention and residual blocks to estimate noise $\hat{\varepsilon}_t$, iteratively refining $\mathbf{z}_t$ to produce the final latent $\mathbf{z}_{\text{final}}$. Training minimizes the Mean Squared Error (MSE) between $\mathbf{z}_{\text{model}}$ and $\mathbf{z}_{\text{true}}$, with the conditional embedding $\mathbf{z}_{\text{cond}}$ guiding model to generate accurate 3D head representation. The loss function for this training process is defined as:

$$L_{\text{total}} = \sum_{t=0}^{T} \lambda_t L(t) \tag{2}$$

where λ_t is a timestep-specific weighting factor, and $L(t)$ represents the MSE loss at timestep t.

3.3 3D Reconstruction via MonoNPHM Decoder

Once we obtain the denoised latent from diffusion, $\mathbf{z}_{\text{final}}$, it is split back into its individual subcomponents $\mathbf{z}_{\text{geo}}$, $\mathbf{z}_{\text{exp}}$, and $\mathbf{z}_{\text{app}}$, and scaled for the decoder (see Section 4.1.1). These act as input to the MonoNPHM decoder that converts these latents to 3D mesh. The decoder combines these latents to generate:

Geometry: SDF(x) to represent the 3D geometry of the head, including both static structural features and dynamic expressions, and

Appearance: RGB(x) to add realistic colour to the reconstructed face.

This provides a high-quality 3D head, preserving identity, expressions, and appearance from the input image.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

4 TRAINING AND INFERENCE

4.1 Dataset

For our training, we used the NPHM dataset, comprising over 9,200 high-quality head scans from 488 identities. Captured with a custom 3D scanning setup, the dataset features diverse facial shapes and expressions, enabling the model to learn detailed and generalized representations of human faces.

4.1.1 Latent Representation of 3D Meshes

To encode essential attributes of facial geometry, expression, and appearance, each 3D mesh is converted into a compact latent representation using the MonoNPHM decoder. This process yields three types of latent codes:

Geometric Latent ($\mathbf{z}_{\text{geo}} \in \mathbb{R}^{2176}$): Encodes the structural features of the 3D face,

Appearance Latent ($\mathbf{z}_{\text{app}} \in \mathbb{R}^{2176}$): Captures the surface colour, and

Expression Latent ($\mathbf{z}_{\text{exp}} \in \mathbb{R}^{100}$): Captures changes in facial morphology.

The latent codes are obtained by initialising random values for each 3D head and then iteratively optimising these initial values. The optimization process involves passing the random latent vectors through the frozen MonoNPHM decoder and minimizing loss. For calculating the loss we precompute SDF(x) and RGB(x) values for near surface points x and optimize for latent codes:

$$\underset{z_{\text{geo}},z_{\text{exp}},z_{\text{app}}}{\operatorname{argmin}} \sum_{x\in P} \lambda_{\text{SDF}} \left|\mathcal{D}_{\text{geo}}(\mathcal{D}_{\text{exp}}(x)) - \text{SDF}(x)\right| + \lambda_{\text{RGB}} \left|\mathcal{D}_{\text{app}}(\mathcal{D}_{\text{geo}}(\mathcal{D}_{\text{exp}}(x))) - \text{RGB}(x)\right| \qquad (3)$$

After optimization, the latent codes are rescaled from an arbitrary range to $[-1,1]$ for diffusion training stability. For example, the geometric latent is rescaled as:

$$z_{\text{geo\_rescaled}} = 2 \times \left(\frac{z_{\text{geo}} - z_{\text{geo\_min}}}{z_{\text{geo\_max}} - z_{\text{geo\_min}}} \right) - 1 \qquad (4)$$

where: $z_{\text{geo\_min}}, z_{\text{geo\_max}}$ are dataset specific extrema. The appearance and expression latents are rescaled similarly.

The rescaled latents are concatenated into a unified vector:

$$\mathbf{z}_{\text{true}} = [z_{\text{geo\_rescaled}}, z_{\text{app\_rescaled}}, z_{\text{exp\_rescaled}}, 0_4] \qquad (5)$$

where $\mathbf{0}_4$ is a zero-padding vector to ensure a fixed dimensionality of 4456.

This final latent vector serves as the ground truth for training the diffusion model, encapsulating detailed structural and expressive features for accurate 3D head representation.

4.1.2 Rendering Input Images from 3D Meshes

To generate inputs for the latent diffusion process, high-quality 3D meshes are rendered into 2D images using Trimesh (Dawson-Haggerty et al.,) and Pyrender (Matl et al., 2019). A scene is set up with the 3D mesh at the origin, a perspective camera for proper framing, and a point light source for illumination. The resulting images are used for embedding generation and serve as diffusion model conditions.

4.2 Inference Process

During inference, the input image I is processed by FaRL to generate a conditional embedding $\mathbf{z}_{\text{cond}} \in \mathbb{R}^{512\times 1}$, which guides the diffusion model. The model iteratively refines the latent representation to generate $\mathbf{z}_{\text{final}}$, which is decoded by the MonoNPHM decoder to reconstruct the 3D mesh geometry of the input face. The results (Fig. 3) show accurate geometry, contours, and features, with minimal deviation in complex regions like the nose, mouth, and eyes. This demonstrates the model's capability to handle diverse geometries and fine details, supporting realistic 3D facial modelling.

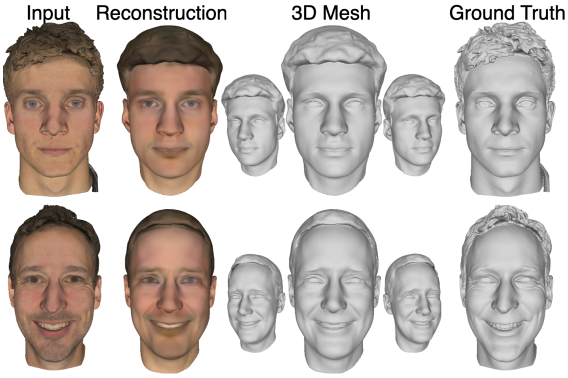


Figure 3: 3D reconstruction comparison against ground truth. Our model accurately reconstructs 3D geometry from a single image, closely matching the ground truth 3D mesh.

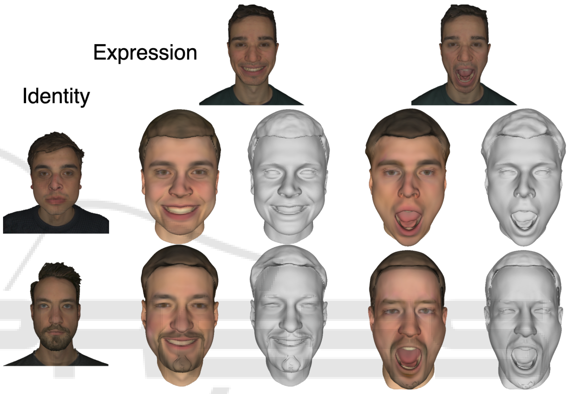


Figure 4: Results of expression transfer. Given an identity image and expression image, transfer of expression on identity.

For the transfer of expression from one image to a different identity, FaRL is used to extract embeddings from both an identity image and an expression image. After the diffusion process, geometry (z_{geo}) and appearance (z_{app}) latents are taken from the identity image, while expression (z_{exp}) is derived from the expression image. These combined latents form the final representation, decoded by the MonoNPHM decoder to produce a 3D head with the first image's identity and the second image's expression. This enables precise manipulation of facial identity and expressions, accurately transferring expressions while preserving identity, as shown in Fig. 4.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

4.3 Implementation Details

We utilise Pytorch (Paszke et al., 2019) for the implementation. For the model architecture, we employed a U-Net (Ronneberger et al., 2015) as the primary diffusion backbone. A single input image of 512 × 512 resolution was used, optimizing data usage while maintaining detailed output fidelity. We use pre-trained FaRL that was trained on LAIONFace20M (Zheng et al., 2022) for 64 epochs. We train our model for 4500 epochs, using a batch size of 24 and a learning rate of $8e-5$ for 48 hours on 2 H100 GPUs.

5 BENCHMARKING AGAINST SOTA

In this section, we present and analyze the results of our model and compare its performance with other SOTA methods in 3D head reconstruction such as Morphable Diffusion (Diffusion-based), HRN (3DMM-based), and MonoNPHM (NPHM-based). We evaluate reconstruction fidelity, focusing on high-precision geometry and expression accuracy, as well as assessing the efficiency of our model in terms of inference time and accuracy-to-time ratio.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

5.1 Evaluation Details

5.1.1 Test Data

We evaluated on the NPHM and Facescape (Zhu et al., 2023a) datasets. For NPHM, a separate test set of six diverse identities was used, including challenging cases like long hair and facial hair. Each subject had a random and neutral expression for testing generalization. On FaceScape, real images of varied subjects were used without retraining, assessing cross-dataset generalization to different settings. This dual evaluation highlights the model's adaptability to new identities and real-world variations.

5.1.2 Metrics

To quantify reconstruction accuracy, we employ the unidirectional L1-Chamfer distance (in centimetres) as our primary evaluation metric.

In addition to accuracy, we evaluate model efficiency by calculating inference time and the accuracy-to-time ratio.

5.1.3 Evaluation Protocol

For a fair comparison, we align each generated mesh with the ground truth using a two-step alignment process: initial manual registration followed by refinement with Iterative Closest Point (ICP) (Besl and McKay, 1992).

Inference time was measured on an NVIDIA RTX 4090 GPU under identical conditions. The evaluation was conducted using the default configurations of each method, with an input image of fixed resolution. Multiple runs were performed to calculate the average inference time, reducing the impact of system variability.

5.2 Quantitative Results

As shown in Table 1, our model achieves the lowest L1-Chamfer distances, with 37.5% and 38.4% improvements for neutral and expressive faces respectively compared to its closest competitor. Evaluations on the FaceScape dataset confirm this trend, with a 41.3% improvement, demonstrating strong generalization Table 2.

Table 1: Quantitative comparison on NPHM test dataset for a neutral and a random expression against SOTA. Reported values are Chamfer distance with average surface error in centimetres.

Accuracy: Neutral Face (↓)				
Model **Subject Ids**	**MonoNPHM**	**HRN**	**Morphable Diffusion**	**Ours**
1	1.28	9.55	1.64	**0.80**
2	0.79	7.91	0.84	**0.38**
3	0.71	8.60	1.35	**0.57**
4	1.69	9.78	2.09	**1.05**
5	1.02	8.95	1.53	**0.67**
6	1.23	9.46	1.88	**0.75**
Average	1.12	9.04	1.56	**0.70**
Accuracy: Expressive Face (↓)				
1 (Lips Down)	1.05	8.12	1.87	**0.68**
2 (Squeeze)	0.89	8.16	1.03	**0.39**
3 (Grin)	0.76	8.17	1.00	**0.76**
4 (Mouth Stretch)	1.97	10.30	2.35	**1.44**
5 (Angry)	1.12	8.21	1.20	**0.49**
6 (Smile)	1.26	8.63	1.62	**0.58**
Average	1.17	8.60	1.51	**0.72**

Table 2: Quantitative comparison on FaceScape dataset against SOTA. Reported values are Chamfer distance with average surface error in centimetres.

Accuracy (↓)				
Model **Subject Ids**	**MonoNPHM**	**HRN**	**Morphable Diffusion**	**Ours**
1	2.55	8.27	1.60	**1.22**
2	1.26	6.21	1.99	**0.88**
Average	1.90	7.24	1.79	**1.05**

Table 3: Comparison against SOTA based on inference time and efficiency (accuracy/time ratio).

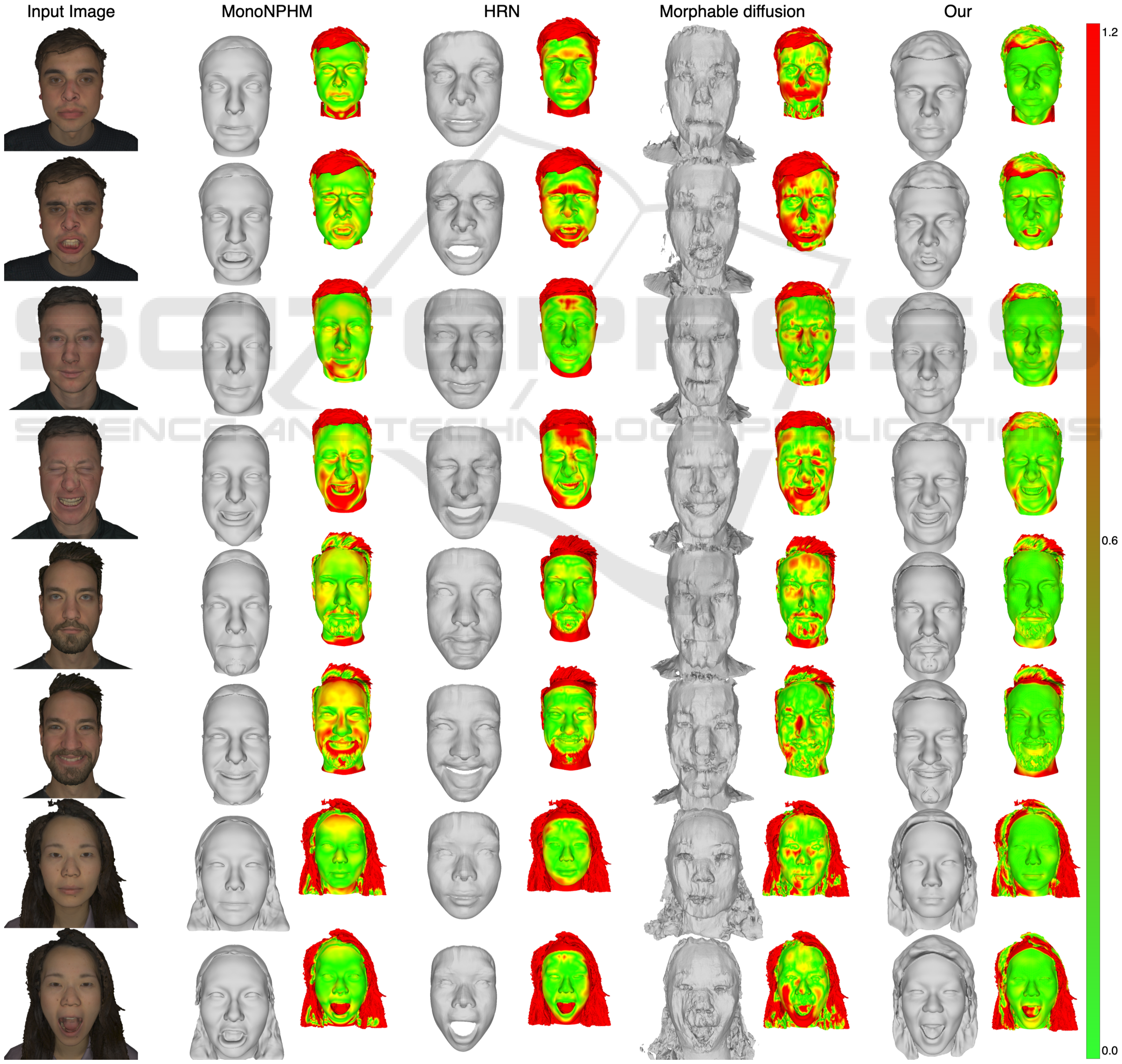
Method	MonoNPHM	HRN	Morphable Diffusion	Ours
Time(mm:ss)	06:55	**00:48**	02:51	1:11
Accuracy/Time	0.0021	0.0023	0.0038	**0.0197**

Table 3 highlights our model's superior balance of speed and accuracy. Unlike multi-step methods like Morphable Diffusion and MonoNPHM, our approach operates directly without added overhead. While HRN is faster, it sacrifices accuracy. Our method's efficiency and performance make it ideal for practical applications.

5.3 Qualitative Results

Fig. 5 provides a qualitative comparison of reconstructed meshes and their L1-Chamfer distances across methods. MonoNPHM generates facial meshes that resemble human faces but often fail to preserve identity and detailed expressions, resulting in higher Chamfer distances for expressive images. HRN better captures expressions but exhibits rough reconstructions and lacks accuracy in modelling facial hair. Morphable Diffusion introduces high-frequency noise, reducing the fidelity of its 3D meshes.

Our model, however, consistently preserves high-fidelity geometry and subtle expressions. Simi-



SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

Figure 5: Qualitative results on NPHM test dataset against MonoNPHM, HRN, and Morphable Diffusion. Odds rows are neutral and even rows are with expression. The colour code is in centimetres.

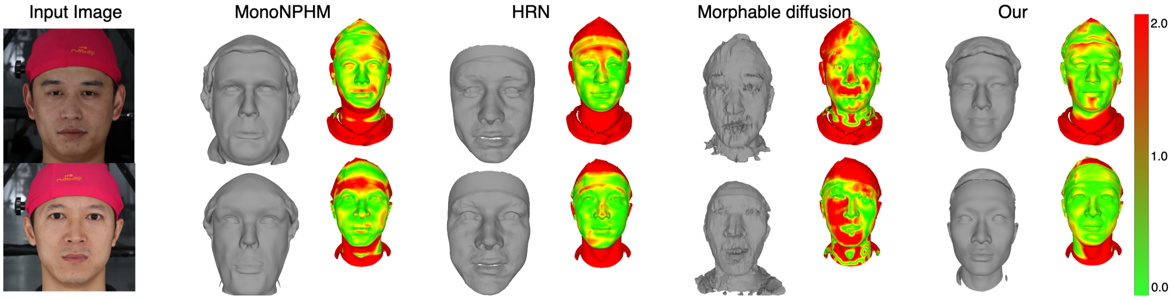


Figure 6: Qualitative results on FaceScape dataset against MonoNPHM, HRN, and Morphable Diffusion. The colour code is in centimetres.

lar trends appear in tests on the FaceScape dataset (Fig. 6), highlighting its robustness in capturing diverse facial structures and expressions under varying conditions.

6 LIMITATIONS

Our model, while strong in 3D head reconstruction and expression transfer, has notable limitations. The use of a 1D latent vector $z_{\text{app}} \in \mathbb{R}^{2176}$ constrains detail and realism of texture. Relying on vertex colours without photorealistic textures or material properties like specular reflections. A modular approach could decouple appearance encoding from geometry and expression, enabling photorealistic texture generation with dedicated models.

Finally, hair information is insufficiently captured in the latent representation or conditional embeddings, leading to hairstyle reconstruction inaccuracies. Specialized mechanisms or hair-specific conditioning inputs could address this gap.

7 CONCLUSION

In this work, we have introduced a 1D diffusion-based method for high-fidelity 3D head reconstruction, capable of generating detailed 3D meshes from a single input image. By leveraging 1D diffusion framework combined with robust conditioning, our approach achieves accurate reconstruction of facial geometry and expressions. The proposed method ensures high-quality structural fidelity and expression accuracy, offering a versatile and scalable solution for realistic 3D head reconstruction. Through comprehensive quantitative and qualitative evaluations, we demonstrate that our method consistently outperforms state-of-the-art approaches, excelling in capturing geometric details and expressive facial variations. This positions our framework as a powerful tool for applications that demand precise and detailed 3D head modelling. Our work marks a significant advancement in 3D head reconstruction, showcasing the potential of small diffusion-based models for generating high-quality 3D meshes. By bridging the gap between accuracy and adaptability, this approach lays the foundation for future innovations.

ACKNOWLEDGEMENT

This work was co-funded by the European Union under Horizon Europe, grant number 101092889, project SHARESPACE. We sincerely thank Dr. René Schuster for their invaluable assistance in writing and proofreading this paper.

REFERENCES

Besl, P. and McKay, N. D. (1992). A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256.

Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194.

Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., and Zafeiriou, S. (2017). 3d face morphable models" in-the-wild". In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 48–57.

Booth, J., Roussos, A., Zafeiriou, S., Ponniah, A., and Dunaway, D. (2016). A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5543–5552.

Bénédict, G., Jeunen, O., Papa, S., Bhargav, S., Odijk, D., and de Rijke, M. (2023). Recfusion: A binomial diffusion process for 1d data for recommendation.

Campbell, N. D., Vogiatzis, G., Hernández, C., and Cipolla, R. (2008). Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Computer Vision–ECCV 2008: 10th European Conference on*

Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10, pages 766–779. Springer.

Cao, C., Weng, Y., Zhou, S., Tong, Y., and Zhou, K. (2013). Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425.

Chai, Z., Zhang, T., He, T., Tan, X., Baltrušaitis, T., Wu, H., Li, R., Zhao, S., Yuan, C., and Bian, J. (2023). Hiface: High-fidelity 3d face reconstruction by learning static and dynamic details.

Dawson-Haggerty et al. trimesh.

Feng, Y., Wu, F., Shao, X., Wang, Y., and Zhou, X. (2021). Deca: Deep face model with 3d morphable models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7668–7677.

Furukawa, Y. and Ponce, J. (2009). Accurate, dense, and robust multiview stereopsis. *IEEE transactions on pattern analysis and machine intelligence*, 32(8):1362–1376.

Giebenhain, S., Kirschstein, T., Georgopoulos, M., Rünz, M., Agapito, L., and Nießner, M. (2023). Learning neural parametric head models. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Grassal, P.-W., Prinzler, M., Leistner, T., Rother, C., Nießner, M., and Thies, J. (2022). Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664.

Gu, J., Gao, Q., Zhai, S., Chen, B., Liu, L., and Susskind, J. (2023). Control3diff: Learning controllable 3d diffusion models from single-view images.

Kirschstein, T., Giebenhain, S., and Nießner, M. (2023). Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. *arXiv preprint arXiv:2311.18635*.

Lei, B., Ren, J., Feng, M., Cui, M., and Xie, X. (2023). A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images.

Li, T., Bolkart, T., Black, M. J., Li, H., and Romero, J. (2017). Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1.

Li, Y., Zhang, Y., Chen, W., Zheng, Y., Li, H., and Yu, J. (2023). Mono-nphm: Monocular neural parametric head model for high-fidelity 3d head reconstruction. *arXiv preprint arXiv:2312.06740*.

Matl, M. et al. (2019). Pyrender. *GitHub Repository*.

Müller, N., Siddiqui, Y., Porzi, L., Bulò, S. R., Kontschieder, P., and Nießner, M. (2023). Diffrf: Rendering-guided 3d radiance field diffusion.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ploumpis, S., Wang, H., Pears, N., Smith, W. A., and Zafeiriou, S. (2019). Combining 3d morphable models: A large scale face-and-head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10934–10943.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. (2022). Dreamfusion: Text-to-3d using 2d diffusion.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation.

Schönberger, J. L., Zheng, E., Frahm, J.-M., and Pollefeys, M. (2016). Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer.

Tran, L., Liu, F., and Liu, X. (2019). Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135.

Tran, L. and Liu, X. (2018). Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355.

Wang, Y., Hu, Y., Guo, Y., Zhang, Z., and Yu, J. (2024). Morphable diffusion: 3d face reconstruction from a single image. *arXiv preprint arXiv:2401.04728*.

Zhao, Y., Liu, Z., Yang, M., and Chen, L. (2022). Rodin: A realistic 3d face reconstruction method using diffusion models. *International Journal of Computer Vision*, 130(3):707–724.

Zhao, Y., Liu, Z., Yang, M., and Chen, L. (2023). Rodinhd: High-resolution 3d face reconstruction using diffusion models. *arXiv preprint arXiv:2305.05555*.

Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., and Wen, F. (2021). General facial representation learning in a visual-linguistic manner. *arXiv preprint arXiv:2112.03109*.

Zheng, Y., Yang, H., Zhang, T., Bao, J., Chen, D., Huang, Y., Yuan, L., Chen, D., Zeng, M., and Wen, F. (2022). General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709.

Zheng Ding, Cecilia Zhang, Z. X. L. J. Z. T. and Zhang, X. (2023). Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhu, H., Yang, H., Guo, L., Zhang, Y., Wang, Y., Huang, M., Wu, M., Shen, Q., Yang, R., and Cao, X. (2023a). Facescape: 3d facial dataset and benchmark for single-view 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Zhu, Y., Ye, Y., Zhang, S., Zhao, X., and Yu, J. J. Q. (2023b). Difftraj: Generating gps trajectory with diffusion probabilistic model.

Zielonka, W., Bolkart, T., and Thies, J. (2022). Towards metrical reconstruction of human faces.