
Proceedings Paper:
Rowan, William, Huber, Patrik, Pears, N. E. et al. (2024) N Heads Are Better Than One: Exploring Theoretical Performance Bounds of 3D Face Reconstruction Methods. In: European Conference on Computer Vision Workshop (ECCVw) 2024:Foundation Models for 3D Humans. Lecture Notes in Computer Science. Springer Science + Business Media, pp. 427-435.

https://doi.org/10.1007/978-3-031-92591-7_28

N Heads Are Better Than One: Exploring Theoretical Performance Bounds of 3D Face Reconstruction Methods

Will Rowan[1], Patrik Huber[1], Nick Pears[1], and Andrew Keeling[2]

[1] University of York, York, United Kingdom
{will.rowan, patrik.huber, nick.pears}@york.ac.uk
[2] University of Leeds, Leeds, United Kingdom. a.j.keeling@leeds.ac.uk

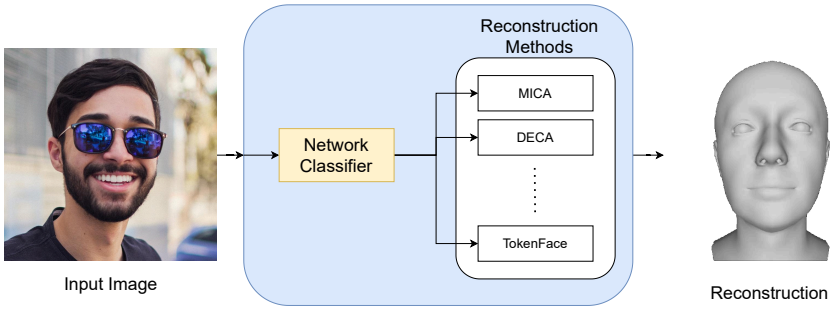

Fig. 1: We evaluate multiple existing 3D face reconstruction methods (such as MICA, DECA, and TokenFace) on a given input image. Our theoretical classifier selects the best reconstruction for each input, enabling us to calculate lower error bounds for combinations of methods and establish new baselines for 3D face reconstruction performance.

Abstract. We introduce "N Heads Are Better Than One", a novel approach for evaluating combinations of existing 3D face reconstruction methods. By calculating lower theoretical error bounds for method combinations on the NoW benchmark, we establish a robust set of new baselines for the task of 3D face reconstruction. Our work also provides a framework for assessing the potential of these aggregate 'pseudo-foundation models,' which leverage strengths from multiple existing approaches. In doing so, we improve understanding of the performance of current methods and set targets for future foundation models to beat.

Keywords: 3D Face Reconstruction · Foundation Models · Performance Evaluation

1 Introduction

3D face reconstruction from 2D images is a fundamental task in computer vision with wide-ranging applications [20]. Despite advancements, accurately esti-

mating 3D face shape from a single image remains an ill-posed problem due to inherent ambiguities [3]. Statistical priors like 3D Morphable Models provide additional constraints, but their use is limited by the size and diversity of available training data [6, 9].

The emergence of foundation models offers potential for improved reconstruction accuracy and user interaction by leveraging vast amounts of data across multiple modalities (image, video, text, 3D). However, realising their full potential requires establishing new, strong, interpretable baselines and developing better analytical tools to inform the design of new methods.

We introduce "N Heads Are Better Than One", a method to calculate lower theoretical error bounds for combining existing 3D face reconstruction methods. This approach provides a new set of strong baselines, demonstrates potential improvements by leveraging complementary strengths of existing methods, and in doing so, develops a range of **pseudo-foundation models** that act as aggregates of existing methods. Our work contributes to the development of a comprehensive foundation model for the human head and expands the understanding of existing 3D face reconstruction methods.

2 Related Work

2.1 3D Face Reconstruction

3D face reconstruction is a long-standing research topic in computer vision. Approaches are broadly categorised into model-based and model-free methods [20]. Model-based approaches, particularly those using 3D Morphable Models (3DMMs) [4], are widely adopted due to their ability to incorporate prior knowledge about face shape and appearance.

The 3D Morphable Model, introduced by Blanz and Vetter [4], is a learned statistical model of the human face that considers both shape and texture. It is constructed by performing dimensionality reduction on a set of training meshes put into dense point-to-point correspondence. This model has become a widely-popular and useful tool for both generation and analysis of the human face, offering a compact representation of facial geometry and appearance. This enables reconstruction to be formulated as an optimisation or regression problem to this learned lower-dimensional space of the human head.

2.2 Deep Learning Approaches and Datasets

Recent advancements in deep learning have significantly improved 3D face reconstruction. Tewari et al. [17] introduced MoFA, a self-supervised approach combining a convolutional encoder with a differentiable renderer, allowing end-to-end training on 2D images alone. Zielonka et al. [19] achieved state-of-the-art results with MICA, leveraging a unified collection of existing 3D face datasets for supervised training. This demonstrates the importance of diverse, high-quality training data for improved reconstruction accuracy, particularly for metric accuracy.

Zhang et al. [18] employed a Vision Transformer for encoding facial components in their TokenFace method, achieving state-of-the-art results on the NoW benchmark [15]. They use a hybrid training strategy, combining supervised training on unified 3D face datasets with self-supervised training on large image datasets. The scarcity of available 3D data presents a key challenge for accurate metric reconstruction, limiting the application of supervised learning in this field. This suggests that further sources of information across multiple modalities (images, text, video, 3D) will be required to improve 3D face reconstruction, an opportunity that foundation models may address.

2.3 Foundation Models in 3D Face Reconstruction

The emergence of foundation models like CLIP [12] and Stable Diffusion [13] has opened new possibilities for 3D face reconstruction. These models, pre-trained on vast datasets, offer rich representations for various downstream tasks. Aneja et al. [1] demonstrated CLIP's potential for text-based editing of 3D face models, enabling intuitive manipulation of facial attributes through natural language. Rowan et al. [14] utilised conditioned Stable Diffusion to generate large-scale paired datasets for 3D reconstruction, enabling supervised learning at scale, though the datasets remain susceptible to generation artifacts. Early attempts at face-specific foundation models include Arc2Face [11], which fine-tunes Stable Diffusion to generate face images from identity descriptors, showing impressive identity consistency but lacking 3D shape grounding. 4M [10] and 4M-21 [2] use expert models to train any-to-any foundation models for generation and analysis, including human poses, shape, depth, and normals. These approaches offer insights into opportunities for a face-specific foundation model. Our method of calculating lower error bounds for combining existing reconstruction techniques provides a framework to evaluate the potential performance of pseudo-foundation models that leverages the strengths of multiple existing approaches.

3 Methodology and Experimental Setup

3.1 N Heads Are Better Than One

It is a common saying that 'two heads are better than one'. This expression captures the intuition that two people considering the same problem are often better than one. We extend this idea to 'N heads are better than one' for 3D face reconstruction. We propose that having N reconstruction methods, each producing a single reconstruction for a given input image, offers the opportunity to develop a challenging set of new baselines for 3D face reconstruction. This can be used both to inform the design of foundation models for digital humans, by better understanding existing methods for face reconstruction, and to assess their performance against the strong new set of baselines our method provides.

3.2 Combining Reconstruction Methods

We aim to determine the theoretical performance achievable by optimally selecting the best reconstruction network for a given image. This can be approached in two ways: as a classification problem to be learned or as an optimisation problem using known error values from a benchmark. We consider the latter case in this paper. Our theoretical approach simulates an idealised classifier that always selects the best network for a given image. This classifier, $C(I_f, S)$, selects the reconstruction network f_r from set S that minimises the error function E for an input image I_f:

$$C(f_i, S) = \underset{f_r \in S}{\operatorname{argmin}} E(I_f, f_r) \tag{1}$$

Where f_i is the input image, S is the subset of reconstruction networks, f_r is a reconstruction network in S, and $E(I_f, f_r)$ is the error function. This approach provides an upper bound on the performance achievable by any practical method selection strategy, serving as a benchmark for evaluating real-world implementations and guiding future research.

3.3 The NoW Benchmark

We implement an idealised classifier as defined in Eq. (1) and evaluate our approach using the NoW benchmark [15], which has become the standard for assessing 3D face reconstruction from 2D images. NoW comprises high-quality images and 3D head scans of 100 subjects (20 validation, 80 test), featuring various poses, occlusions, and expressions. Using the test set, we calculate theoretical lower error bounds for combining reconstruction methods across metric and non-metric reconstruction. We include all publicly available methods on the NoW benchmark, except 3DFFA-V2 due to incomplete error data.

4 Results and Analysis

4.1 Performance of Combined Methods

To quantify the potential of method combination and provide a comprehensive analysis of progress in the field, we present the results of combining various existing methods in Table 1. This table includes combinations grouped by year of publication (pseudo-foundation models), as well as novel combinations we introduce, such as DICA (DECA + MICA) and TICA (TokenFace + MICA).

Our analysis reveals several significant findings:

1. **Progress over time:** The time-grouped results, visualised in Figures 2a and 2b, demonstrate a clear trend of improvement in both metric and non-metric reconstruction. The combination of all methods up to 2023 consistently outperforms individual methods by a substantial margin, indicating the potential benefits of combining multiple approaches.

Method	Non-Metric			Metric (mm)		
	Median	Mean	Std	Median	Mean	Std
FLAME mean [9]	1.21	1.53	1.31	1.49	1.92	1.68
Deng et al. (PyTorch) [5]	1.23	1.54	1.29	2.26	2.90	2.51
RingNet [15]	1.21	1.53	1.31	1.50	1.98	1.77
MGCNet [16]	1.31	1.87	2.63	1.70	2.47	3.02
DECA [7]	1.09	1.38	1.18	1.35	1.80	1.64
MICA [19]	0.90	1.11	0.92	1.08	1.37	1.17
FOCUS [8]	1.04	1.30	1.10	1.41	1.85	1.70
TokenFace [18]	0.76	0.95	0.82	0.97	1.24	1.07
DICA	0.87	1.09	0.92	1.04	1.34	1.17
TICA	0.75	0.94	0.82	0.93	1.20	1.05
FICA	0.86	1.08	0.92	1.04	1.33	1.17
TECA	0.75	0.95	0.83	0.95	1.22	1.07
TOCUS	0.75	0.95	0.83	0.95	1.21	1.07
2018 & earlier	1.42	1.85	1.73	3.91	4.84	4.02
2019 & earlier	0.99	1.27	1.12	1.37	1.86	1.72
2020 & earlier	0.97	1.26	1.19	1.26	1.72	1.65
2021 & earlier	0.94	1.23	1.15	1.14	1.51	1.39
2022 & earlier	0.80	1.02	0.93	0.93	1.20	1.06
2023 & earlier	0.72	0.93	0.84	0.86	1.12	0.99

Table 1: Results for both metric and non-metric reconstruction on the test set of the NoW benchmark. We compile results for existing methods and compare them with theoretical lower bounds for combinations of methods, both by date of publication and for novel combinations of existing methods such as DICA which is the lower error bound for combining MICA and DECA.

2. **TokenFace performance:** Notably, TokenFace [18] surpasses the combination of all pre-2022 methods in non-metric reconstruction. This significant improvement in shape recovery accuracy can be attributed to its novel use of facial component tokens and vision transformers.
3. **Complementary method strengths:** The TICA combination (TokenFace + MICA) exhibits improved performance in metric reconstruction compared to individual methods, while offering minimal improvements in non-metric reconstruction over TokenFace alone. This suggests that MICA contributes additional strength in recovering facial scale, while TokenFace excels in capturing precise facial shapes.
4. **Diminishing marginal returns:** While we observe continuous improvement, the relative gains appear to be diminishing, particularly in non-metric reconstruction. This trend may indicate an approach towards the limits of current datasets and training methods, emphasising the need for new architectures or the integration of additional data modalities.
5. **Differential progress in metrics:** Our analysis reveals that metric reconstruction has experienced more substantial relative improvements since the introduction of the NoW benchmark. This suggests that recent meth-

ods have been particularly effective in addressing challenges related to facial scale over precise facial details.

These findings demonstrate the potential benefits of developing hybrid approaches that leverage the strengths of multiple existing methods to achieve more accurate 3D face reconstructions. They also provide valuable insights into the current state of the field, indicating promising directions for future research.

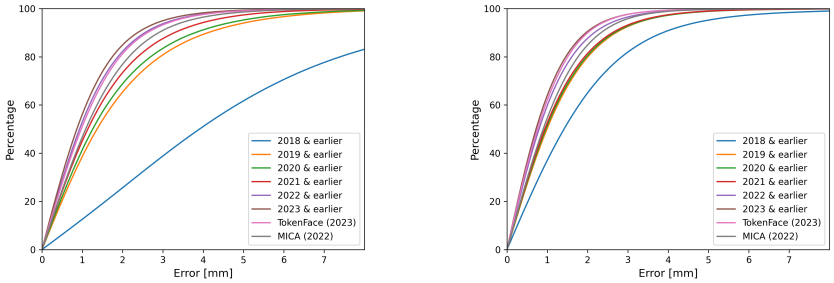

Fig. 2: Error plots for the NoW benchmark in metric and non-metric reconstruction. We compare errors from two leading approaches and the combination of all methods grouped by their year of publication. 18 and 14 methods are considered respectively.

4.2 Error Reduction Over Time

Table 2 presents a quantitative assessment of error reductions over time for both metric and non-metric errors in 3D face reconstruction on the NoW benchmark. We analyse the errors for combined methods grouped by year of release. For example, '2022 to 2023' represents the performance difference between theoretically combining all methods released up to 2022 and those released up to 2023.

Our analysis reveals improvements in both metric and non-metric reconstruction over time. Notably, metric reconstruction has experienced the largest relative improvement since the introduction of the NoW benchmark. Specifically, we observe a cumulative error reduction of 77.96% for metric reconstruction, compared to 48.72% for non-metric reconstruction between 2018 and 2023.

It is important to note that this comparison is made at an aggregate level of methods available at certain points in time, rather than a direct comparison of individual methods. This is because several methods only report results for non-metric reconstruction, as they are designed to recover facial details but not facial scale.

Year	Median (%)	Mean (%)	Std (%)	Cumulative (%)
Non-Metric Errors				
2018 to 2019	30.28	31.35	35.26	30.28
2019 to 2020	2.02	0.79	-6.25	31.66
2020 to 2021	3.09	2.38	3.36	33.77
2021 to 2022	14.89	17.07	19.13	43.91
2022 to 2023	10.00	8.82	9.68	48.72
Metric Errors				
2018 to 2019	64.96	61.57	57.21	64.96
2019 to 2020	8.03	7.53	4.07	67.48
2020 to 2021	9.52	12.21	15.76	70.81
2021 to 2022	18.42	20.53	23.74	76.21
2022 to 2023	7.53	6.67	6.60	77.96

Table 2: Percentage decrease in errors over time for non-metric and metric errors. Cumulative reductions are calculated for the median errors reported.

5 Conclusions and Future Work

We have presented *N Heads Are Better Than One*, a novel approach that establishes new baselines for 3D face reconstruction by calculating lower theoretical error bounds for combining existing methods. Our results demonstrate the significant potential for improvement in this field by leveraging the complementary strengths of existing methods.

Our work makes several key contributions. We provide new, strong baselines that are readily communicable, improving our understanding of existing methods and offering new targets for future methods to beat. We identify the gap between current state-of-the-art methods and the optimal theoretical performance achieved through combining existing methods, revealing the potential for substantial improvement in 3D face reconstruction techniques. Furthermore, we demonstrate the marked progress in the field over time, particularly in metric reconstruction accuracy.

These findings pave the way for future advancements, particularly in the development of foundation models for 3D face reconstruction. Such models could play a crucial role in more effectively leveraging the strengths of multiple approaches. Future work should aim to implement practical methods for combining existing reconstruction techniques. This could involve developing a method classifier trained to select the best approach for a given input image or combining methods at the feature level. Our lower error bounds can serve as benchmarks for evaluating the performance of these new combined approaches.

In conclusion, our work contributes to a better understanding of the strengths and limitations of existing 3D face reconstruction techniques. By offering a range of new targets on existing benchmarks, we provide clear goals for future research in this field, which will offer targets for a foundation model of the human face to beat.

References

1. Aneja, D., Sanyal, S., Ghosh, P.: Clipface: Text-guided editing of textured 3d morphable face models. In: 2022 International Conference on 3D Vision (3DV). pp. 1194–1203. IEEE (2022) 3
2. Bachmann, R., Kar, O.F., Mizrahi, D., Garjani, A., Gao, M., Griffiths, D., Hu, J., Dehghan, A., Zamir, A.: 4m-21: An any-to-any vision model for tens of tasks and modalities. arXiv preprint arXiv:2406.09406 (2024) 3
3. Bas, A., Smith, W.A.: Does face recognition work for everyone? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 2
4. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. Proceedings of the 26th annual conference on Computer graphics and interactive techniques pp. 187–194 (1999) 2
5. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: 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. pp. 0–0 (2019) 5
6. Egger, B., Smith, W.A., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al.: 3d morphable face models—past, present, and future. ACM Transactions on Graphics (TOG) **39**(5), 1–38 (2020) 2
7. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (ToG) **40**(4), 1–13 (2021) 5
8. Li, C., Morel-Forster, A., Vetter, T., Egger, B., Kortylewski, A.: Robust model-based face reconstruction through weakly-supervised outlier segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 372–381 (2023) 5
9. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017), https://doi.org/10.1145/3130800.3130813 2, 5
10. Mizrahi, D., Bachmann, R., Kar, O., Yeo, T., Gao, M., Dehghan, A., Zamir, A.: 4m: Massively multimodal masked modeling. Advances in Neural Information Processing Systems **36** (2024) 3
11. Papantoniou, F.P., Lattas, A., Moschoglou, S., Deng, J., Kainz, B., Zafeiriou, S.: Arc2face: A foundation model of human faces. arXiv preprint arXiv:2403.11641 (2024) 3
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021) 3
13. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 3
14. Rowan, W., Huber, P., Pears, N., Keeling, A.: Fake it without making it: Conditioned face generation for accurate 3d face shape estimation. arXiv preprint arXiv:2307.13639 (2023) 3
15. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7763–7772 (2019) 3, 4, 5
16. Shang, J., Shen, T., Li, S., Zhou, L., Zhen, M., Fang, T., Quan, L.: Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In: European Conference on Computer Vision. pp. 53–70. Springer (2020) 5
17. Tewari, A., Zollhöfer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE international conference on computer vision. pp. 1274–1283 (2017) 2
18. Zhang, Z., Chen, Y., Jiang, Y., Sun, X., Zhou, H., Dai, Q., Zhang, L.: Accurate 3d face reconstruction with facial component tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18148–18157 (2023) 3, 5
19. Zielonka, W., Bolkart, T., Thies, J.: Towards metrical reconstruction of human faces. In: European Conference on Computer Vision. pp. 730–746. Springer (2022) 2, 5
20. Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., Theobalt, C.: State of the art on monocular 3d face reconstruction, tracking, and applications. Computer Graphics Forum **37**(2), 523–550 (2018) 1, 2