# TEXT2FACE: A MULTI-MODAL 3D FACE MODEL

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We present the first 3D morphable modelling approach, whereby 3D face shape can be directly and completely defined using a textual prompt. Building on work in multi-modal learning, we extend the FLAME head model to a common image-and-text latent space. This allows for direct 3D Morphable Model (3DMM) parameter generation and therefore shape manipulation from textual descriptions. Our method, Text2Face, has many applications; for example: generating police photofits where the input is already in natural language. It further enables multi-modal 3DMM image fitting to sketches and sculptures, as well as images.

## 1 INTRODUCTION

Generative 3D shape models, such as 3D Morphable Models (3DMMs) (Blanz & Vetter, 1999), are useful statistical priors with which to explain 2D/3D images of 3D objects, such as human faces, by reconstructing their shape. This has enabled applications in personalised avatar design (Lombardi et al., 2018), medical diagnosis of fetal alcohol syndrome (Suttie et al., 2013), and prosthesis design for missing facial regions (Mueller et al., 2011). Furthermore, there are a multitude of generative applications. For example, Blanz et al. (2006) generate police photofits from eyewitness accounts, but they rely on manual editing of 3DMM parameters. Instead, we propose a method which allows descriptive text to be directly mapped to the latent space of the 3DMM. In doing so, among other applications, we enable photofits to be initialised directly from witnesses' textual descriptions or a sketch.

Creating 3D faces from text prompts paves the way for further shape refinement using suitable textual prompts which drive major/minor shape adjustments. Furthermore, given the availability of an initial textual description, the method enables a better (than average) shape initialisation in applications which employ model-to-image fitting.

CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) is a pre-trained visual-textual embedding model. This approach has shown strong zero-shot capabilities on a wide range of computer vision tasks and has enabled models to take advantage of this embedding space for downstream tasks, such as text-to-image generation (Ramesh et al., 2021). Several approaches have used the expressive power of CLIP to relate text to 3D shape. This includes text-driven generation of stylised meshes (Michel et al., 2022); general text to shape generation (Sanghi et al., 2022); text-based texture and expression editing (Aneja et al., 2022); and full body 3D avatar creation and animation (Hong et al., 2022). However, none of these methods consider the generation of a fully parameterised model of the human face, including identity, such as a 3DMM.

In this work, we bring together 3DMMs and CLIP to enable text to 3DMM generation. To do this, we train a deep MLP, Text2Face, to map from CLIP embedding space to the space of 3DMMs and show the strong qualitative results of this approach.

## 2 PROPOSED METHOD

Our method enables us to generate a fully paramaterised 3D model of the human head, including identity, expression, and a detail map, from a single text prompt. To do this, we first generate a dataset of mappings between CLIP embedding space and the parameter space of the FLAME model.

We synthesise 50,000 adult faces using StyleGan2 (Karras et al., 2020), selecting images estimated to be older than 18 using py-agender (Butlitsky, 2018). A CLIP embedding is extracted from each

image using the ViT-L/14-336px vision transformer model (Radford et al., 2021). We further estimate identity, pose, and expression vectors for each image in FLAME model space (Li et al., 2017) using DECA (Feng et al., 2021), a state of the art method for monocular 3D face reconstruction. We extract identity, expression, pose, and a detailed displacement map, $\delta$, for each image.

We train a deep MLP, Text2Face, on this dataset to map the CLIP embedding space to the FLAME parameter space. At inference time, we take advantage of CLIP's interchangeable text-image latent space by using CLIP's text encoder on a sequence of input text. We supply this CLIP embedding as input to Text2Face which generates parameters for a fully paramaterised 3D face. These are passed to the DECA decoder (Feng et al., 2021) to produce the 3D mesh. Hence, training Text2Face on embeddings extracted from images allows for inference for both text and images. We further use DALL-E (Ramesh et al., 2021) to generate an image for each text prompt, using texture mapping to map this to the mesh. The overall architecture is shown in Figure 1.

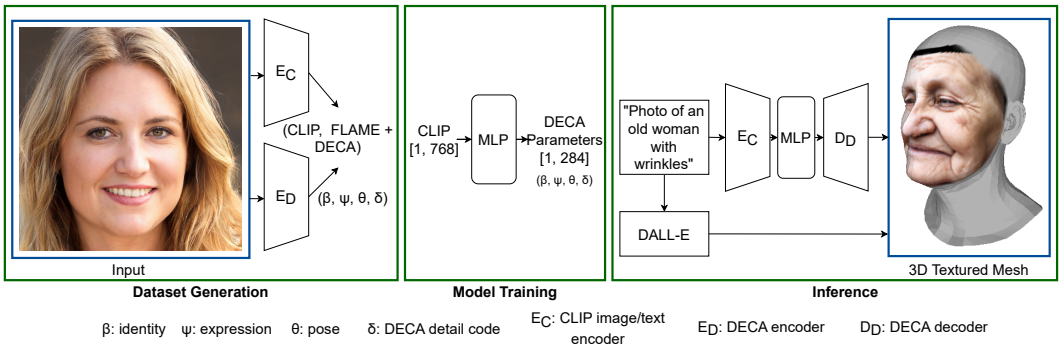


Figure 1: Dataset generation, model training, and inference for Text2Face.

We use the Adam optimiser (Kingma & Ba, 2014) with a learning rate of 1e-3 and a batch size of 64. We train for 100 epochs, using early stopping with a patience of 10. A full network diagram is presented in Appendix B (Figure 11).

## 3 Experiments

Figure 1 shows the resulting textured 3D mesh from the text prompt: "Photo of an old woman with wrinkles". Figure 2 shows the detailed texture-less mesh generated by the text prompt. The image generated by DALL-E from this same prompt and subsequent textured mesh are also shown. Further qualitative results are shown in Appendix A.

Figure 2: *(l)*: Prompt: "20 year old woman looking at the sky with surprise at UFO overhead". *(r)*: Prompt: "50 year old man looking grumpy". Each sub-figure, from left to right: Shape generated by the text prompt, DALL-E image from same prompt, textured mesh.

## 4 Conclusion

We have presented the first method for text to fully parameterised 3D face shape. This finds application in photofit specification, avatar creation, and wider 3DMM fitting settings. Further work should consider inherited gender and racial biases from CLIP Agarwal et al. (2021), their impact on 3D face generation, and how this can be minimised.

## URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

## REFERENCES

Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

Shivangi Aneja, Justus Thies, Angela Dai, and Matthias Nießner. Clipface: Text-guided editing of textured 3d morphable models. *arXiv preprint arXiv:2212.01406*, 2022.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pp. 187–194, 1999.

Volker Blanz, Irene Albrecht, Jörg Haber, and H-P Seidel. Creating face models from vague mental images. In *Computer Graphics Forum*, volume 25, pp. 645–654. Wiley Online Library, 2006.

Michael Butlitsky. py-agender, 2018. URL `https://github.com/`.

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018.

Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13492–13502, 2022.

AA Mueller, Pascal Paysan, Ralf Schumacher, H-F Zeilhofer, B-I Berg-Boerner, Juerg Maurer, Thomas Vetter, Erik Schkommodau, Philipp Juergens, and Katja Schwenzer-Zimmerer. Missing facial parts computed by a morphable model and transferred directly to a polyamide laser-sintered prosthesis: an innovation study. *British journal of oral and maxillofacial surgery*, 49(8):e67–e71, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. Clip-forge: Towards zero-shot text-to-shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18603–18613, 2022.

Michael Suttie, Tatiana Foroud, Leah Wetherill, Joseph L Jacobson, Christopher D Molteno, Ernesta M Meintjes, H Eugene Hoyme, Nathaniel Khaole, Luther K Robinson, Edward P Riley, et al. Facial dysmorphism across the fetal alcohol spectrum. *Pediatrics*, 131(3):e779–e788, 2013.

## A Further Qualitative Evaluation

In the following figures 3 to 9, we show three images. The first image shows the 3D mesh constructed from the 3DMM parameters regressed by Text2Face from the specified text prompt. The second image is an image generated by Ramesh et al. (2021) from the same text prompt. The final image shows this generated image mapped onto the generated mesh as texture. This full pipeline is implemented to enable a textured 3D mesh to be generated directly from a single text prompt.

The pose of the mesh displayed in the first image is fit to the generated DALL-E image. The identity, expression, and detail code are unchanged, all being regressed directly by Text2Face from the text prompt.

Figure 3: Prompt: ”Happy man”.

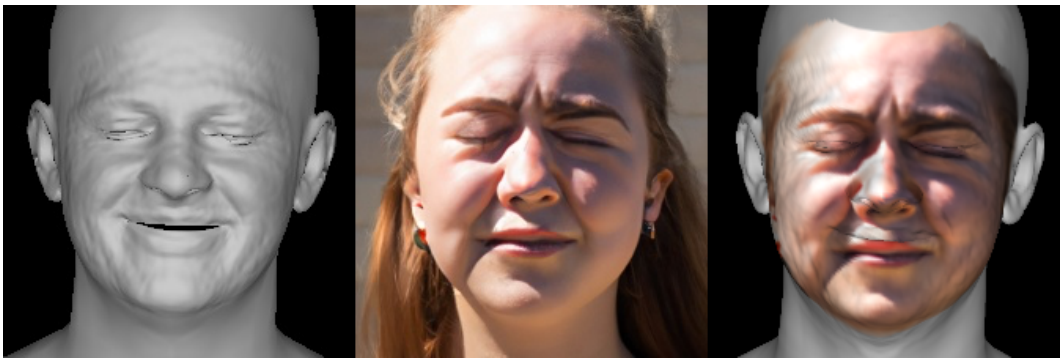

Figure 4: Prompt: ”20 year old woman squinting at the sun”.

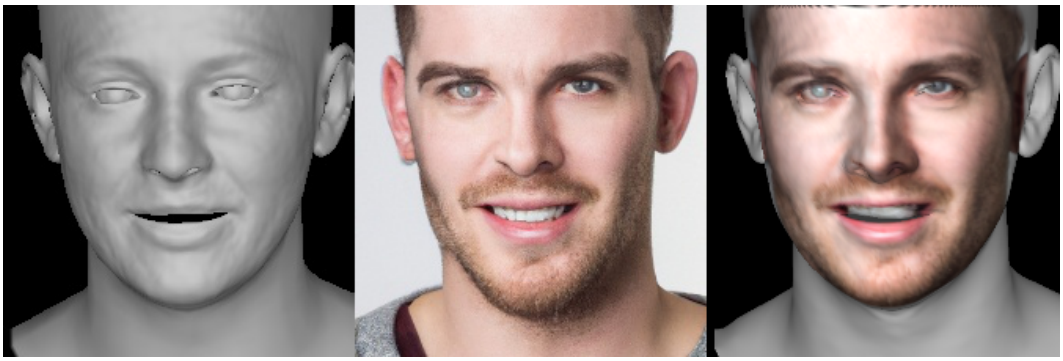

Figure 5: Prompt: ”24 year old attractive man”.

Figure 6: Prompt: ”Photo of a woman screaming”.

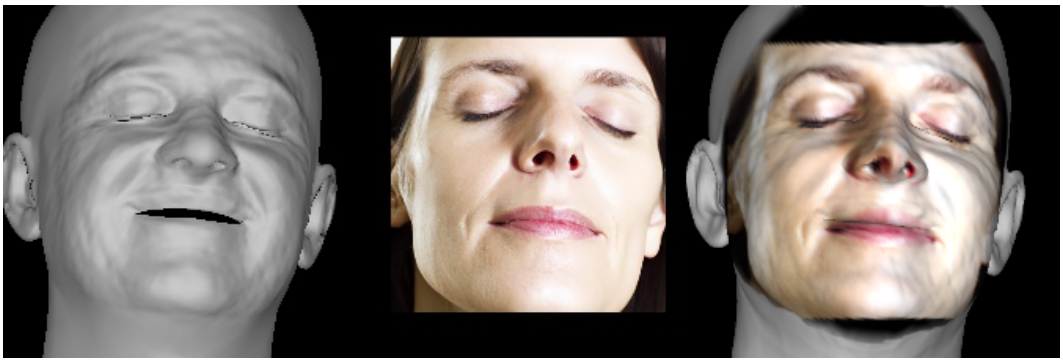

Figure 7: Prompt: ”Photo of a woman with her eyes closed”.

Figure 8: Prompt: "50 year old man looking happy after a long day working on the film set".

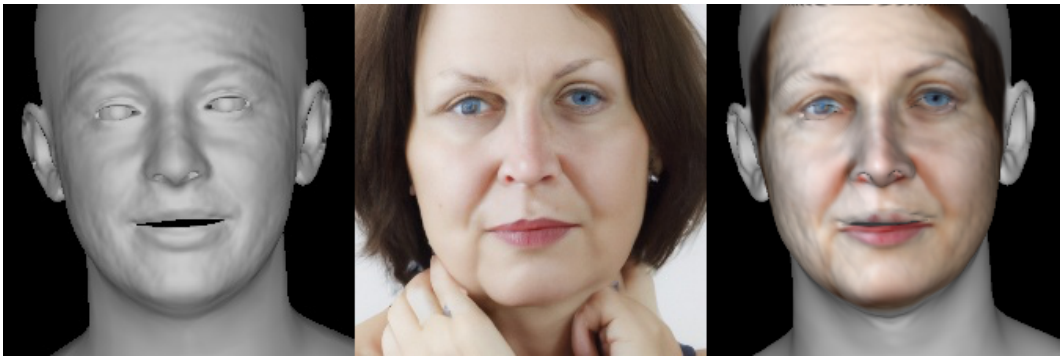

Figure 9: Prompt: "50 year old woman".

### A.1 MULTI-MODAL FITTING

Here we further demonstrate the ability of Text2Face to enable multi-modal input fitting to a 3DMM. We show this through the use of different types of images. We take an image of Robert DeNiro and create sketch and sculpture versions using image processing techniques. We extract the CLIP embedding from each image using the ViT-L/14-336px vision transformer model Radford et al. (2021) and pass this as input to Text2Face which regresses the 3DMM paramaters, including identity, expression, and a personal detail code.

The result is shown in Figure 10. We present all three images in a pose aligned with the original image of DeNiro. We also show texture mapping results for all generated meshes using the first image to generate this texture.

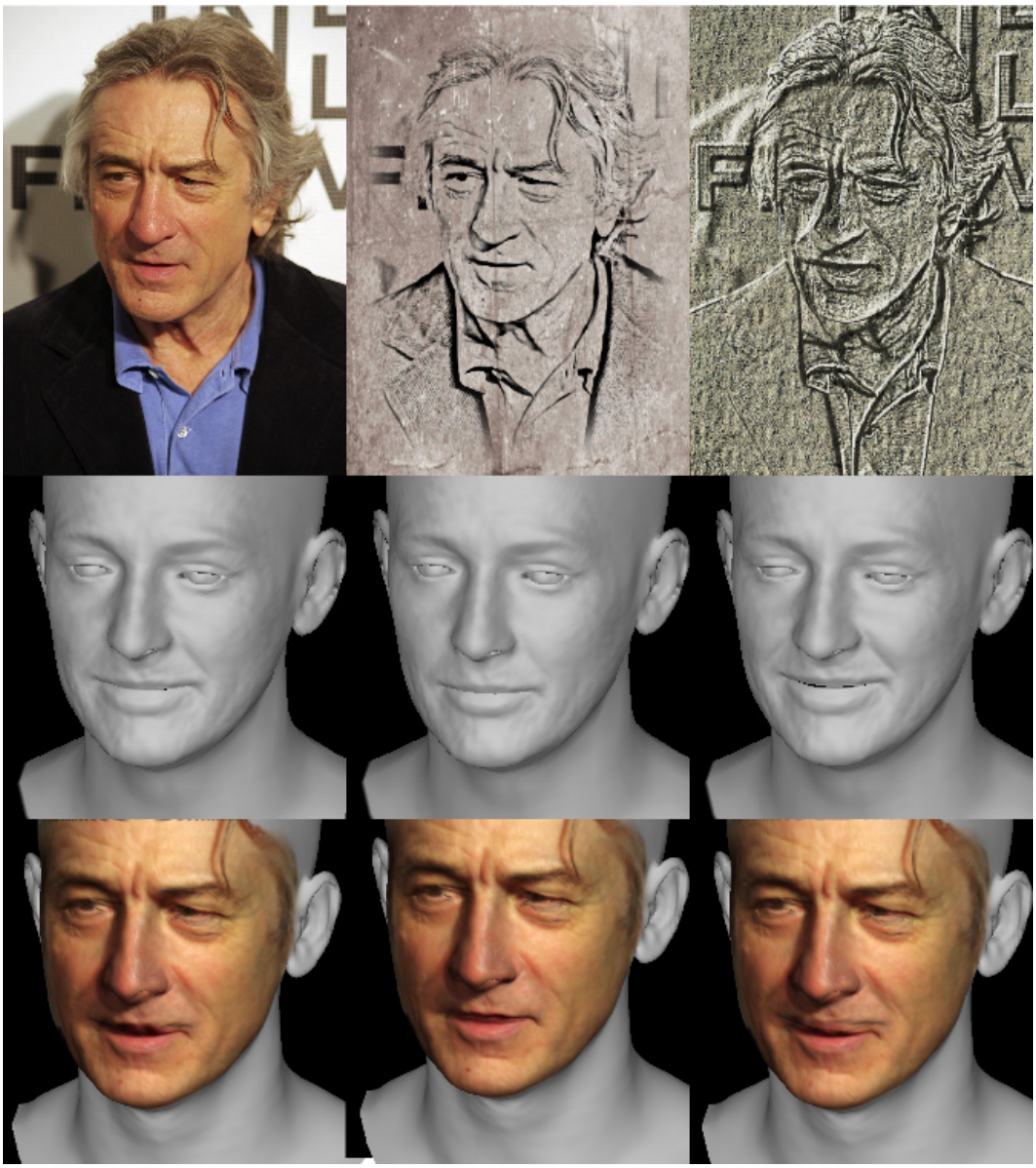

Figure 10: Robert DeNiro fit to a 3DMM using Text2Face: using an original image (left), a sketch (middle), and an engraving (right).

## B TEXT2FACE ARCHITECTURE

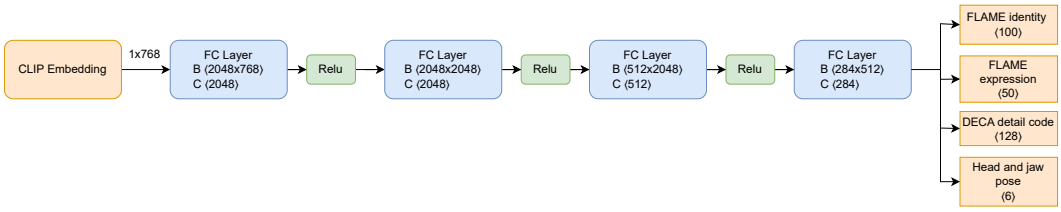


Figure 11: Architectural diagram of our Text2Face regressor.