ScanTalk: 3D Talking Heads from Unregistered Scans. Supplementary Material

Federico Nocentini*[1], Thomas Besnier*[2], Claudio Ferrari[4], Sylvain Arguillere[5], Stefano Berretti[1], and Mohamed Daoudi[2,3]

[1] Media Integration and Communication Center (MICC), University of Florence, Italy
federico.nocentini@unifi.it, stefano.berretti@unifi.it
[2] Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
thomas.besnier@univ-lille.fr
[3] IMT Nord Europe, Institut Mines-Télécom, Centre for Digital Systems
mohamed.daoudi@imt-nord-europe.fr
[4] Department of Architecture and Engineering University of Parma, Italy
claudio.ferrari2@unipr.it
[5] Univ. Lille, CNRS, UMR 8524 Laboratoire Paul Painlevé, Lille, F-59000, France
sylvain.arguillere@univ-lille.fr

In this supplementary material, we provide additional details and results that did not fit into the main paper.

A Ethical Comments

We recognize the ethical considerations surrounding the creation of 3D facial animations. Generating synthetic narratives with 3D faces poses inherent risks, potentially resulting in both intentional and unintentional consequences for individuals and society as a whole. We emphasize the importance of adopting a human-centered approach in the development and implementation of such technology.

Human-centered design is essential for shaping technology-driven strategies that benefit humans. Our goal in this work is to develop technology that helps people and addresses an open problem in the literature. Speech-driven human face animation has numerous applications, some of which may be beneficial, while others could be negative. Here we emphasize the importance of responsible use, relying on the end user to apply the technology we developed properly.

B Transformer Decoder

Inspired by prior works [1,3], ScanTalk with Transformer Decoder follows a distinct approach. The architecture is shown in Fig. 1 and it employs a *SpeechEncoder* module preceding an autoregressive Transformer Decoder, which necessitates an initial token. Unlike traditional methods, our approach initializes the

* Equal contribution

generation process with the global representation of $m_i^{neutral}$, the neutral face for animation, denoted as g_i^n, serving as the starting token. The per-vertex features are aggregated through averaging, yielding:

$$g_i^n = \frac{1}{V_i} \sum_{k=1}^{V_i} (f_i^n)_k \in \mathbb{R}^h. \tag{1}$$

This global feature vector, g_i^n, encapsulates fundamental attributes of the neutral face, providing valuable insights into its overall structure and characteristics. While Faceformer [1] commences generation with an embedding of a one-hot label representing the speaker, and Imitator [3] begins from a zero token, our methodology offers a novel perspective on initializing the generation process.

The Transformer Decoder comprises a concatenation of components: a *Positional Encoding Layer* encoding token positions in the sequence, a *Masked Self-Attention Layer* incorporating information from preceding tokens, and a *Masked Cross-Attention Layer* combining token information with corresponding details from the *SpeechEncoder*. The autoregressive token generation process is defined as:

$$v_i^j = TD(v_i^{1:j-1}, a_i^j) \in \mathbb{R}^h \qquad \forall j = 1, \ldots, T_i \quad with \quad v_i^0 = g_i^n. \tag{2}$$

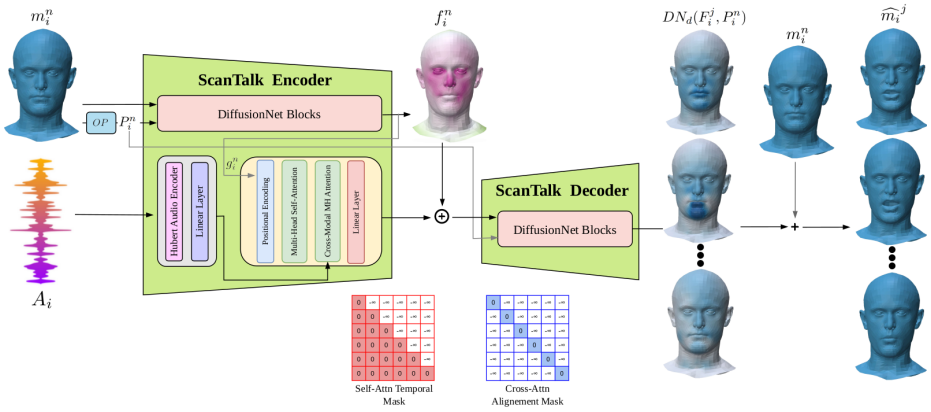

Fig. 1: Architecture of **ScanTalk** Transformer.

C Mesh encoding

Several encoding strategies for geometry are feasible; however, with our experimentation we saw that encoding vertex positions provides the optimal and most

intuitive approach. When omitting mesh encoding and directly feeding the BiLSTM output to the decoder, the mesh signal remains constant across frames, leading to a static facial expression as the decoder lacks spatial awareness of the mouth's location. Incorporating normals alongside positions fails to enhance results, as precomputed operators already furnish adequate orientation information. Additionally, adopting the Heat Kernel Signature (HKS), as suggested in the DiffusionNet framework, does not yield improvements in results. In Fig. 2, we present the per-vertex norm of features derived from the DiffusioNet Encoder for both training and testing meshes.

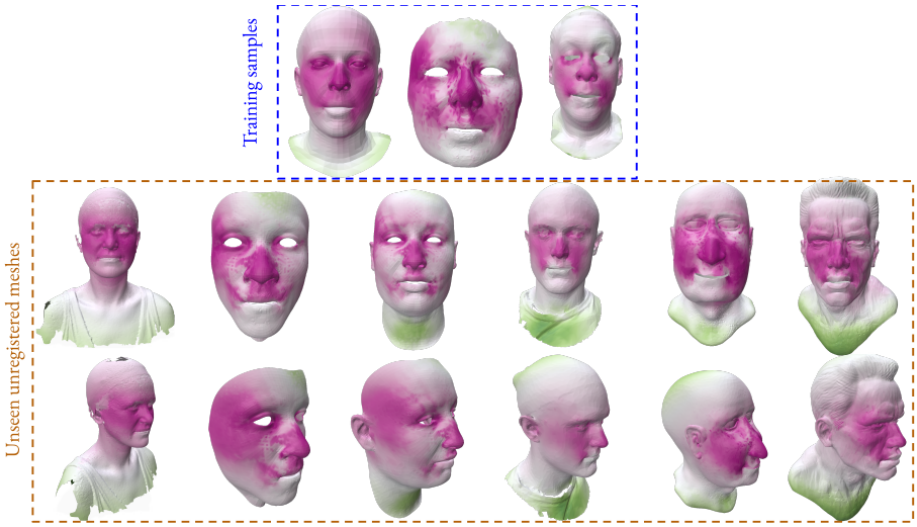

Fig. 2: Relative norm of the per-vertex descriptors extracted by the encoder displayed as a heatmap where pinker hues indicates lower values and greener hues indicates higher values.

D Datasets

We summarize the characteristics of the datasets in Tab. 1. Our preprocessing of the BIWI dataset is depicted in Fig. 3, while the manipulation applied to the Multiface dataset is illustrated in Fig. 4. Specifically, the BIWI dataset underwent downsampling and rigid alignment with the VOCAset, whereas the Multiface dataset was rigidly aligned with the VOCAset, with additional modifications involving the creation of three apertures corresponding to the eyes and mouth.

Table 1: Train / test / val splits for each dataset.

	VOCAset	**BIWI$_6$**	**Multiface**
Type	Head + neck	Narrow face	Head + neck
# vertices	5,023	3,895	5,471
# faces	9,976	7,539	10,837
Training samples	320	400	410
Val samples	80	80	100
Test samples	80	80	100

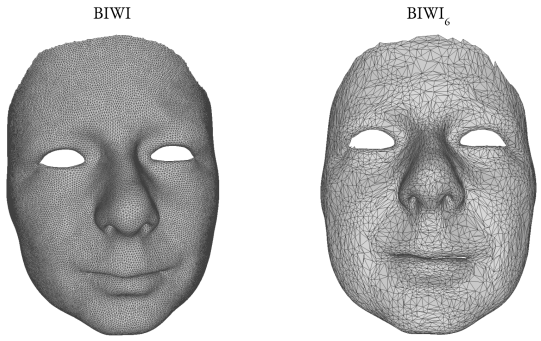

Fig. 3: Side by side comparison of an original mesh from BIWI and the same mesh in BIWI$_6$.

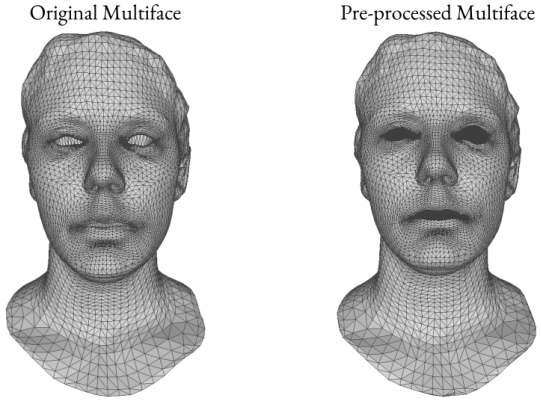

Fig. 4: Side by side comparison of an original mesh from Multiface and the same mesh after preprocessing.

E Additional Qualitative results

In Fig. 5, we present qualitative examples of animation using ScanTalk applied to 3D faces with arbitrary topology. Our preprocessing steps included rigid

alignment with training meshes and the creation of an aperture for the mouth. From Fig. 5 it is evident that ScanTalk exhibits a remarkable capacity for generalization, enabling animation of any 3D face once aligned with the training set and provided with a mouth aperture. Notably, ScanTalk demonstrates effectiveness in animating diverse 3D face meshes, including non-human variants. Such versatility holds significant promise for applications spanning video game development and virtual reality animation.

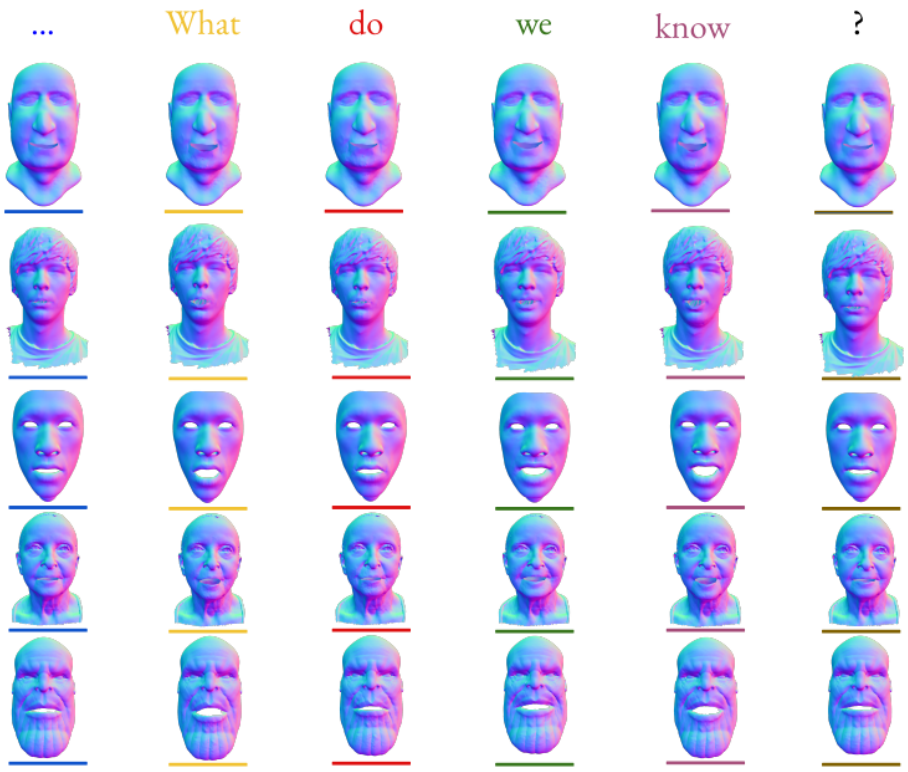

Fig. 5: Additional experiments with different unseen meshes.

F Implementation details

Our ScanTalk model, as described in Section 3 of the main paper, is constructed as follows: the DiffusionNet Encoder comprises 4 DiffusionNet blocks, each with a hidden size (h) of 32. The Bi-LSTM consists of 3 layers with a hidden size of 32. The DiffusionNet Decoder accepts as input the concatenation of features of dimension 64 and outputs the per-vertex deformation of the neutral face.

The DiffusionNet Decoder is composed of 4 DiffusionNet blocks concatenated together.
All the ScanTalk versions presented in the main paper are trained for 200 epochs over each dataset using the Adam optimizer [2], with a learning rate of 10^{-4}.

G ScanTalk Limitations

While ScanTalk exhibits commendable performance in lip motion synthesis, we refrained from incorporating expressions due to the persistently limited availability of such data. Exploring how a similar framework to ScanTalk would handle these additional modalities could be an engaging avenue for future research. Although ScanTalk can animate unregistered meshes, its current training process necessitates meshes with a common topology within each sequence. To overcome this limitation, a fully unsupervised training strategy with a loss function that does not rely on point-wise correspondence, yet captures small geometric displacements, could be a promising direction for further exploration.

H User Study Interface

In Fig. 6, we depict the interface presented to users during our User Study detailed in Section 4.6 of the main paper. On the left, the interface for Test 1, an A/B test, is displayed, while the interface for Test 2 is showcased on the right.

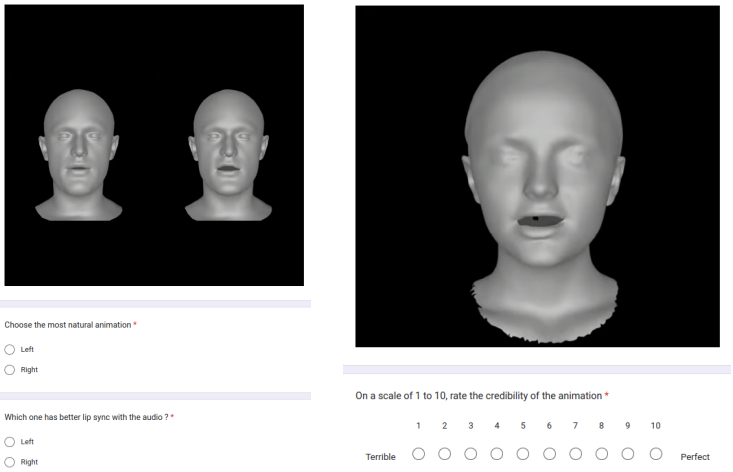

Fig. 6: Examples of questions asked during the user study. (Left) Test 1, an A/B test to compare ScanTalk against state-of-the-art models. (Right) Test 2, we asked the users to evaluate the credibility of scan animations generated by ScanTalk.

References

1. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T.: Faceformer: Speech-driven 3d facial animation with transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). p. 18749–18758. IEEE, New Orleans, LA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01821, https://ieeexplore.ieee.org/document/9878591/
2. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2017)
3. Thambiraja, B., Habibie, I., Aliakbarian, S., Cosker, D., Theobalt, C., Thies, J.: Imitator: Personalized speech-driven 3d facial animation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20621–20631 (October 2023)