EUROGRAPHICS 2022 / R. Chaine and M. H. Kim
(Guest Editors)

Volume 41 (2022), Number 2

Shape Transformers: Topology-Independent 3D Shape Models Using Transformers - Supplemental Material

Prashanth Chandran[1,2] Gaspard Zoss[2] Markus Gross [1,2] Paulo Gotardo[2] Derek Bradley[2]

[1]ETH Zurich [2]DisneyResearch|Studios

1. Ablation

In this section, we evaluate several design choices and measure their effect on the performance of the Shape Transformer. Since training the Shape Transformer exhaustively on a complete dataset, under all possible configurations was a challenge computationally, we used a reasonably sized toy dataset consisting of 10 subjects from the SDFM dataset [CBGB20] in 24 different expressions, resulting in a total of 240 shapes for our ablation study. We primarily analyzed two aspects of performance *(i)* the convergence of the network and *(ii)* the reconstruction quality to make a design choice.

1.1. Choice of Architecture

In this experiment, we evaluate the effect of 4 incremental changes to a naive transformer decoder in a Shape Transformer setting. We begin with *(i)* a naive XCiT transformer decoder with a single linear layer to predict the output offsets, and include our style modulation in variant *(ii)*. Variant *(iii)* adds a transformer encoder to variant *(iii)*. Finally, variant *(iv)* is the full Shape Transformer model that replaces the single layer offset predictor in the previous variants with a 4 layer MLP with residual connection. This effect of these incremental changes on the performance of the Shape Transformer is shown in Fig. 1.

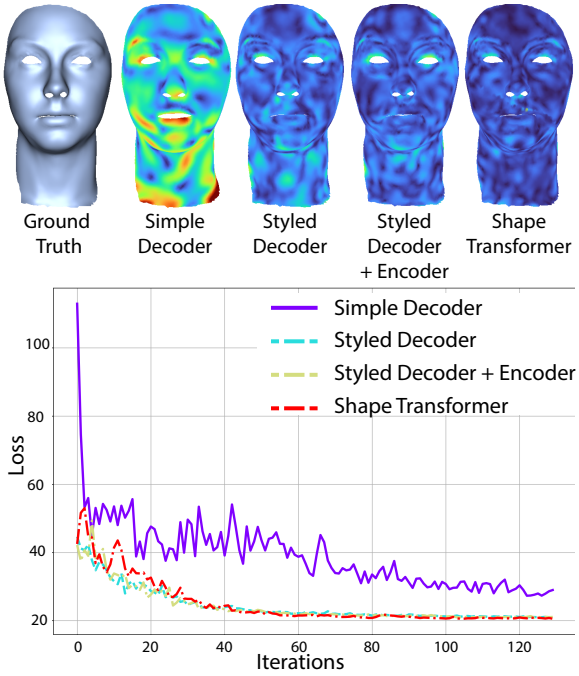

Figure 1: *Our architectural changes incrementally improve the reconstruction quality, with the proposed Shape Transformer architecture achieving an optimal balance between reconstruction quality and convergence.*

1.2. Choice of Attention Mechanism.

Another design choice in the Shape Transformer is the attention mechanism. As we discuss in Section 3.2 of the main paper, Self attention [VSP*17] suffers from quadratic complexities and cannot be used for very long sequences, and thus motivated our use of the cross covariance attention introduced in [ENTC*21]. In Fig. 2, we show the effect of choice of attention on the Shape Transformer. We find that Standard attention achieves better results, but is impractical for our purposes due to its limiting sequence lengths. We note that the Shape Transformer will benefit from any improvements made to the self attention mechanism in transformers as it remains a heavily researched topic.

1.3. Additive vs. Concatenated Position Encoding.

We also evaluate the choice of additive vs. concatenated position encoding. As seen in Fig. 3, concatenating the latent positions and the latent offsets works best in our use case.

1.4. Choice of Activation Function.

We now measure the effect of different activation functions used by the transformer blocks in our architecture, on the convergence and reconstruction quality. We compare three different activation functions, GeLU, ReLU and Sine. Fig. 4 illustrates the reconstruction

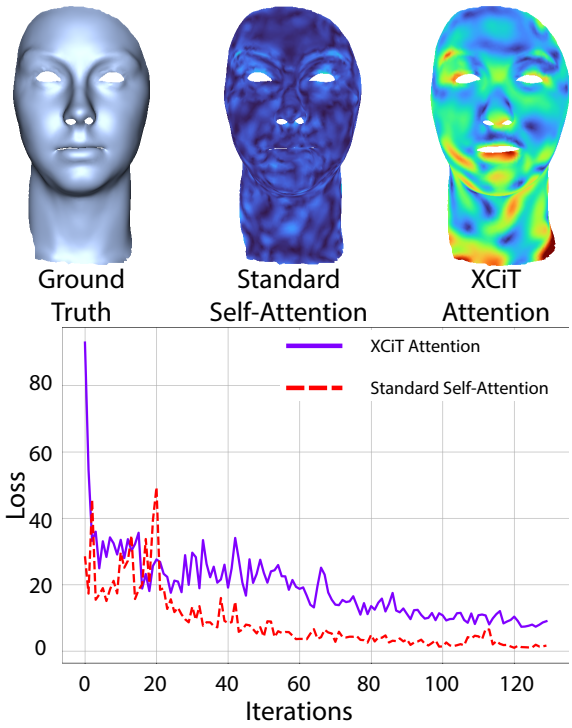

Figure 2: *An ablation study on the attention mechanism used in the transformer decoder*

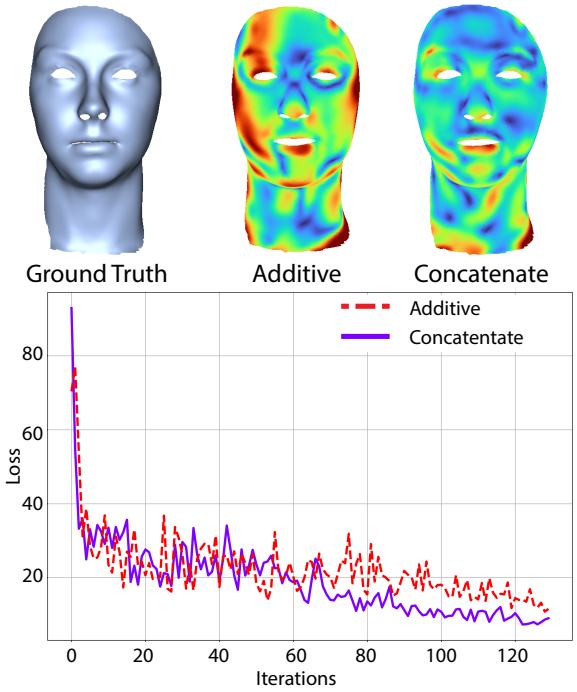

Figure 3: *An ablation study on the position encoding scheme used to train the Shape Transformer indicated concatenated position encoding works best for this setting.*

errors using the different functions, as well as a plot of the convergence. As can be seen, GeLU activation functions provide the best quality and convergence speed.

References

[CBGB20] CHANDRAN P., BRADLEY D., GROSS M., BEELER T.: Semantic deep face models. In *Int. Conf. on 3D Vision* (2020), pp. 345–354. 1

[ENTC*21] EL-NOUBY A., TOUVRON H., CARON M., BOJANOWSKI P., DOUZE M., JOULIN A., LAPTEV I., NEVEROVA N., SYNNAEVE G., VERBEEK J., ET AL.: Xcit: Cross-covariance image transformers. *arXiv preprint arXiv:2106.09681* (2021). 1

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. U., POLOSUKHIN I.: Attention is all you need. In *NeurIPS* (2017), vol. 30. 1

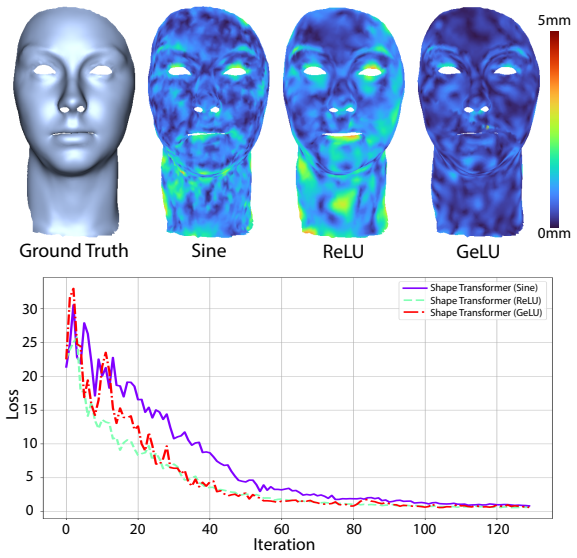

Figure 4: *Shape Transformers are built with GeLU activation functions because they provide the highest reconstruction quality and fast convergence rate.*