# Improving self-supervised 3D face reconstruction with few-shot transfer learning

Martin Dornier
martin.dornier@interdigital.com
InterDigital
Cesson-Sévigné, France

Philippe-Henri Gosselin
philippehenri.gosselin@interdigital.com
InterDigital
Cesson-Sévigné, France

Christian Raymond
christian.raymond@irisa.fr
Univ. Rennes, CNRS, IRISA, France
Rennes, France

Yann Ricquebourg
yann.ricquebourg@irisa.fr
Univ. Rennes, CNRS, IRISA, France
Rennes, France

Bertrand Coüasnon
bertrand.couasnon@irisa.fr
Univ. Rennes, CNRS, IRISA, France
Rennes, France

## ABSTRACT

While self-supervised 3D face reconstruction models have improved over the years, because their training loss is mainly based on the photometric loss, they struggle to predict a 3D face with a correct head pose. On the other hand, supervised methods can predict more accurate head pose but require a lot of annotated data. In this work, we propose to improve self-supervised methods by adding 3D information to their input to improve the predicted head pose. We encode the 3D information in the form of the Projected Normalized Coordinate Code (PNCC). To reduce the need for annotated data to generate the PNCCs, we use transfer learning to adapt a pre-trained face autoencoder to predict the PNCCs. Our PNCC predictor can be trained using only a few annotated samples. Our experiments on a self-supervised method shows that the addition of the PNCC improves the predicted head pose.

## CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**; **Transfer learning**; *Semi-supervised learning settings.*

## KEYWORDS

3D face reconstruction, transfer learning, self-supervised learning

**ACM Reference Format:**
Martin Dornier, Philippe-Henri Gosselin, Christian Raymond, Yann Ricquebourg, and Bertrand Coüasnon. 2023. Improving self-supervised 3D face reconstruction with few-shot transfer learning. In *Proceedings of ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG '23).* ACM, New York, NY, USA, 2 pages.

## 1 INTRODUCTION

3D face reconstruction has many applications, notably in facial animation, but acquiring 3D annotations typically requires a scanner to capture 3D face scans which limits the number of subjects in the dataset. Some datasets try to overcome this limitation by using face model fitting to annotate images but it often leads to poor annotations. Self-supervised methods avoid this issue by training using only 2D face images without 3D annotations, minimizing the photometric error between the original and the reconstructed face images. However, since no 3D information is used during the training, these methods tend to predict a wrong head pose.

We propose to improve the quality of self-supervised methods, especially the predicted head pose, by incorporating face shape and pose information in the architecture, using as few annotated samples as possible. To do so, we use a two-stage process (see Figure 1) composed of a supervised transfer learning stage and a self-supervised stage. The first stage involves adapting a pre-trained face autoencoder to make it predict the Projected Normalized Coordinate Code (PNCC) [Zhu et al. 2016] of an input face image. This modified autoencoder can be trained with only 50 annotated samples. Once trained, we use it to augment, with the predicted PNCCs, a face image dataset. In the second stage, we train a self-supervised 3D face reconstruction model on the augmented dataset but with the PNCC as additional input. Our model predicts better head poses compared to the original model without PNCC.

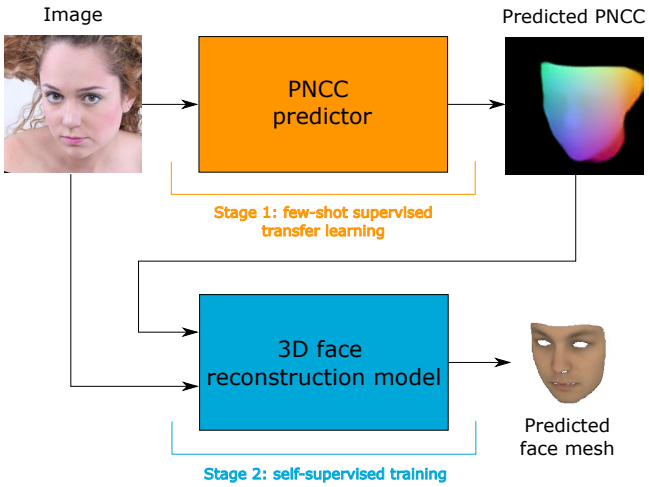


**Figure 1: Our two-stage framework for 3D face reconstruction training.**


*MIG '23, November 15–17, 2023, Rennes, France*

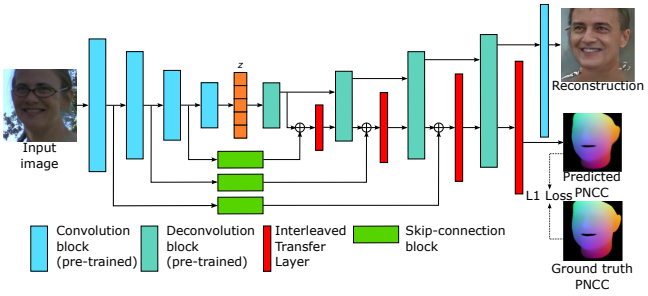


**Figure 2: Our PNCC predictor architecture which can be trained with limited annotated data.**

## 2 METHOD

We train two networks: a PNCC predictor (see Figure 2) and a 3D face reconstruction network. The first one is the pre-trained face autoencoder from 3FabRec with Interleaved Transfer Layers (ITLs) [Browatzki and Wallraven 2020] to make it predict the PNCC. ITLs are convolutional layers interleaved with the decoder layers. They re-use the generative power of the decoder layers but adapt it to generate the PNCC instead of the face image. We also add the skip-connections from SCAF [Dornier et al. 2022] to improve the PNCC quality. Because the encoder and decoder are already pre-trained, only the ITLs and skip-connections layers are trained from scratch. Thus, we are able to train our PNCC predictor with only a few 3D annotations. This training is supervised using samples from 300W-LP [Zhu et al. 2016]. Our self-supervised 3D face reconstruction model is based on MoFa [Tewari et al. 2017] but we stack the PNCC with the face image at the input of its encoder. We train on CelebA [Liu et al. 2015] augmented with our predicted PNCCs.

## 3 EXPERIMENTS

We use 4 evaluation metrics. The *2D Dense Alignment* and *3D Dense Alignment* metrics are the Normalized Mean Error (NME) on the 2D and 3D vertices respectively. The *3D Face Reconstruction* metric first aligns the predicted and ground truth faces meshes prior to computing the NME so it is invariant to the predicted head pose and only evaluates the face shape. We also evaluate the predicted head pose using the Mean Absolute Error (MAE) for the yaw angle. The evaluation is done on AFLW2000-3D [Zhu et al. 2016].

We have trained 2 PNCC predictors to annotate CelebA, one on the whole 300W-LP (122,450 samples): PNCC$_{full}$ and another using only 50 samples: PNCC$_{few}$. We then trained 2 3D face reconstruction models, MoFaPNCC$_{full}$ using PNCC$_{full}$'s predictions and MoFaPNCC$_{few}$ which used PNCC$_{few}$'s predictions. Table 1 reports our results on AFLW2000-3D compared to MoFa. For the dense alignment metrics, both MoFaPNCC$_{full}$ and MoFaPNCC$_{few}$ achieve better results compared to MoFa demonstrating that adding the PNCC to the model input improves the predicted head pose, even when using PNCCs predicted from a model trained with limited annotated data. For the 3D face reconstruction metric the vanilla MoFa is a bit better. However, the main goal of our architecture is to improve the predicted head pose and this metric does not take into account the predicted head pose. The addition of the PNCC also improves the predicted yaw angle although, the gain is quite small for MoFaPNCC$_{few}$. Figure 3 shows the predicted face meshes of MoFa and our models on some images. Our models predict better head pose or face scale.

**Table 1: Evaluation metrics on AFLW2000-3D. MoFaPNCC$_{few}$ uses predictions from our PNCC predictor trained with only 50 samples.**

| Method | Dense 2D | Dense 3D | Face Rec. | Yaw |
|---|---|---|---|---|
| MoFa [Tewari et al. 2017] | 4.31 | 5.85 | 7.49 | 4.97 |
| MoFaPNCC$_{few}$ (Ours) | 4.20 | 5.66 | 7.61 | 4.95 |
| MoFaPNCC$_{full}$ (Ours) | 4.12 | 5.48 | 7.55 | 4.66 |

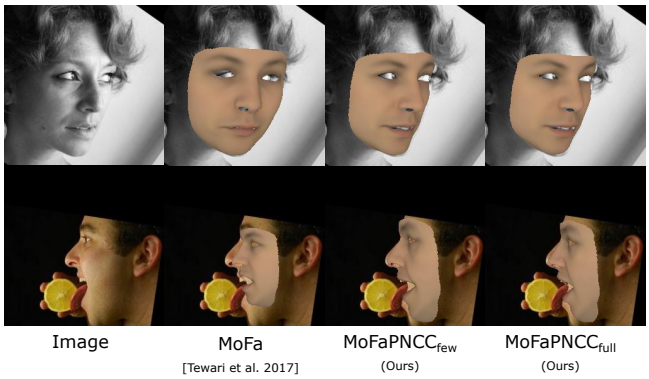


**Figure 3: Comparison of 3D face reconstruction predictions.**

## 4 CONCLUSION

We have proposed an effective way to improve self-supervised 3D face reconstruction methods using only a few 3D face annotations. We used transfer learning to train with limited annotated data a pre-trained face autoencoder to generate PNCCs from face images. Once trained, we used it to annotate a face dataset. We have shown that the PNCCs can be used as additional input to a self-supervised 3D face reconstruction model and improve the predicted head pose.

## ACKNOWLEDGMENTS

This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011012376 made by GENCI.

## REFERENCES

Bjorn Browatzki and Christian Wallraven. 2020. 3fabrec: Fast few-shot face alignment by reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6110–6120.

Martin Dornier, Philippe-Henri Gosselin, Christian Raymond, Yann Ricquebourg, and Bertrand Coüasnon. 2022. SCAF: Skip-Connections in Auto-encoder for Face Alignment with Few Annotated Data. In *International Conference on Image Analysis and Processing*. Springer, 425–437.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision*.

Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1274–1283.

Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. 2016. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 146–155.