# A Decoupled 3D Facial Model by Adversarial Training
# Supplementary Material

## Identity-Expression-Viseme Model

One of the benefits of our framework lies in its ability to easily extend to other factors of variation. As an illustration, we trained a model that decouples identity, expression and viseme (the visual counterpart of a phoneme). The results can be found in Figure 1, where we show qualitative examples obtained by modifying the different factors of variation individually.

We trained the model using the audiovisual 3D dataset of Fanelli *et al.* [1], which contains sequences of $14$ subjects performing $40$ speech sequences in neutral and "expressive" mode. We assign phoneme labels using the Montreal Forced Aligner tool [2] with the provided audio, which are mapped to visemes following [3]. For expression, we manually labeled $699$ frames with the aid of the provided expression ratings of each sequence. This resulted in a database with $100\%$ labeled identites, $68\%$ labeled visemes, and $3\%$ labeled expressions. We set the latent dimensions to $(50, 50, 50, 5)$ for identity, expression, viseme and noise, respectively.

Note this is a simplified model of speech, since the temporal information is not taken into account. Yet, we can see in Figure 1 that a decoupling between the aspects affected by phoneme production, and those affected by expressions such as happiness or surprise can be easily distinguished by our framework. It is also worth noting that these results were obtained with fully automatic labels for viseme, and very sparse manual labels for expression, thus simplifying the efforts required to annotate the dataset. Unlike the identity and expression factors, which are intuitively easier to separate, the viseme and expression factors are more intertwined and decoupling them is very challenging even for a human annotator. In spite of this, our results show that we can reasonably decouple the three factors.

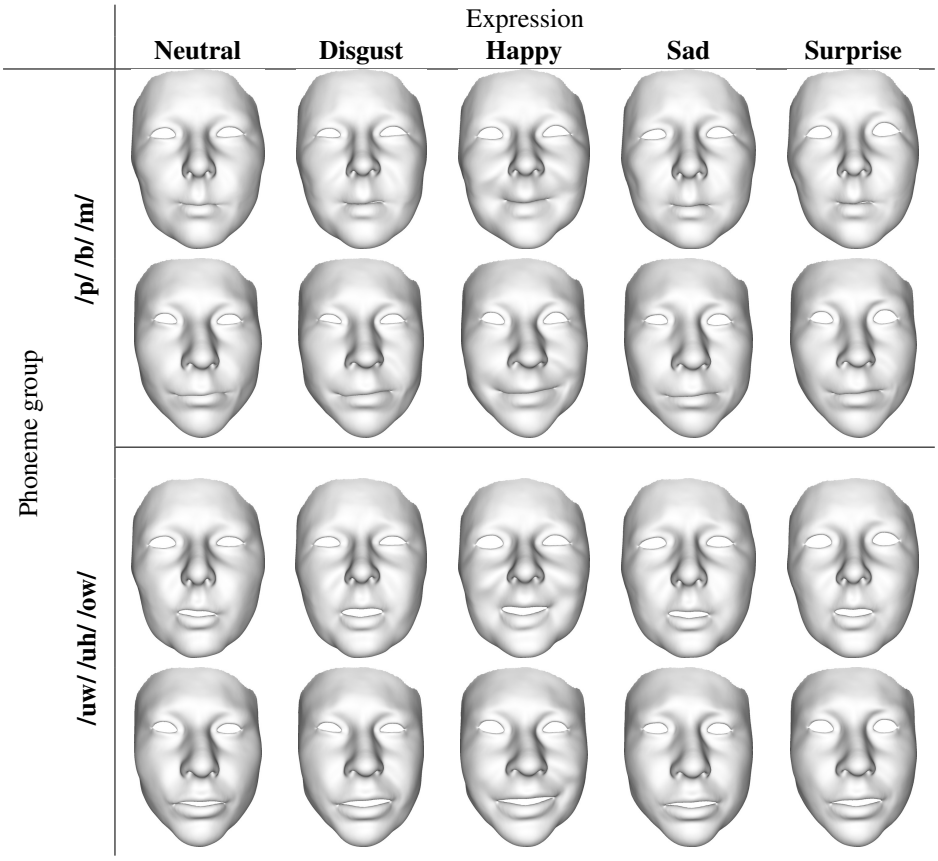



Figure 1: Example of decoupling between identity, expression and viseme.

## Latent Space Manipulation

The following figure shows an example of interpolation and extrapolation in (1) the expression latent space, (2) the identity latent space, and (3) the full latent space:

Figure 2: From top to bottom: interpolation (purple) and extrapolation (gray) of expression code, identity code, and the full latent.

Thanks to the decoupling of identity and expression spaces, we can synthesize new expressions by simple manipulation of the latent space. We show here two possibilities for this.

Given a source mesh obtained with $G(z_{id}^{src}, z_{expr}^{src}, z_{noise}^{src})$ and a target mesh obtained with $G(z_{id}^{target}, z_{expr}^{target}, z_{noise}^{target})$, we generate new expressions for the target mesh by either

1. Replacing the expression with that of the source: $G(z_{id}^{target}, \mathbf{z}_{\mathbf{expr}}^{\mathbf{src}}, z_{noise}^{target})$
2. Adding the expression vectors: $G(z_{id}^{target}, \mathbf{z}_{\mathbf{expr}}^{\mathbf{src}} + \mathbf{z}_{\mathbf{expr}}^{\mathbf{target}}, z_{noise}^{target})$

Results can be seen in Figure 3. In particular, note how adding the latent vectors results in plausible expressions which preserve the semantics of both sources.


Figure 3: Example of expression space manipulation. In gray a source mesh and a target mesh. In blue the result of (1) replacing the expression code of the target with that of the source *(replaced)*, and (2) adding the source and target expression codes *(added)*.

## Qualitative Comparisons

This section provides qualitative examples for the results in Section 5.5, Table 1. Figure 4 shows three random samples with best and worst specificity values, and Figures 5 and 6 show random samples used for decoupling and diversity evaluation of identity and expression, respectively.

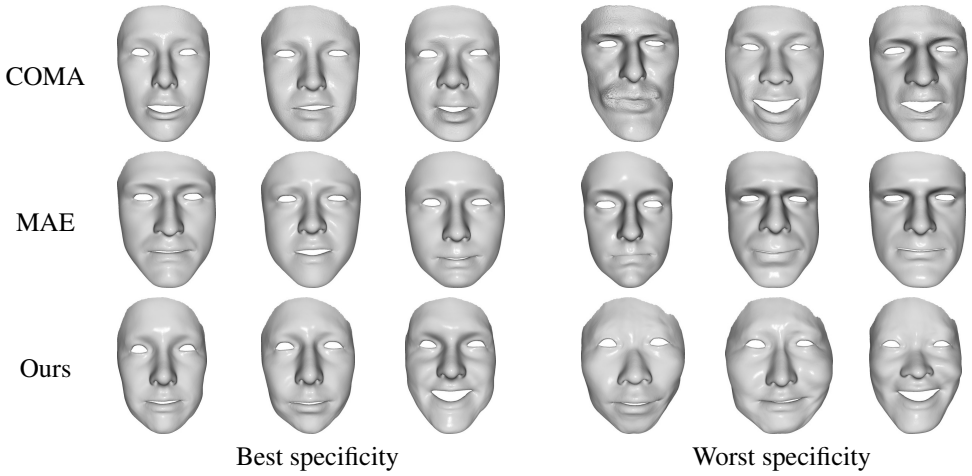



Figure 4: Random samples which obtained the three best (left) and worst (right) values in the specificity metric.

(a) COMA

(b) MAE

(c) Ours

Figure 5: Example of results used for identity decoupling and diversity evaluation, for the three compared methods. Each row shows samples with a same identity code, while the expression code is drawn randomly. Note the low variability in the generated samples for MAE, as also seen in Table 1.

(a) COMA

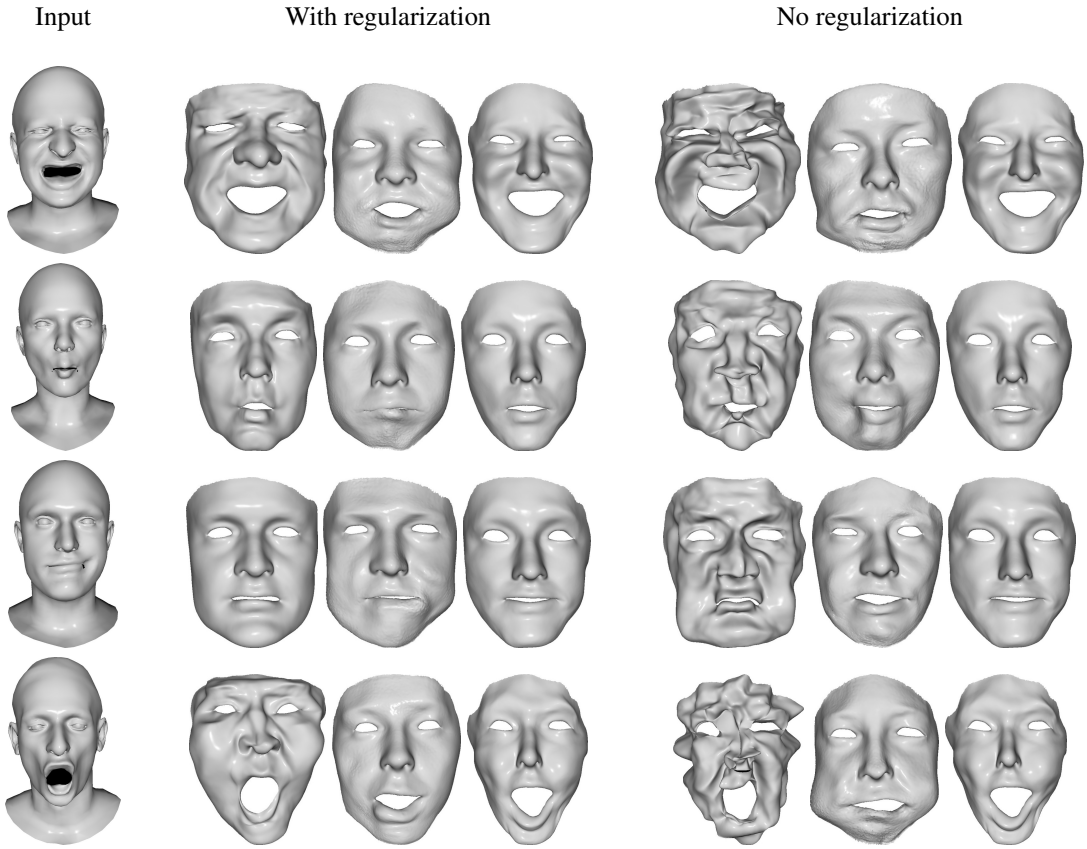
(b) MAE

(c) Ours

Figure 6: Example of results used for expression decoupling and diversity evaluation, for the three compared methods. Each row shows samples with a same expression code, while the identity code is drawn randomly.

## Reconstruction of Sparse Data

Figure 7a shows qualitative results for the experiment in Table 2. The landmarks used for this evaluation are shown in Figure 7b.




(a) Comparison against MAE and COMA, with and without regularization. From left to right: MAE, COMA, our result.

(b) 85 landmarks used for fitting

Figure 7: Reconstruction of sparse data

## Architecture Details

In Figure 8 we show the architecture for the Generator and Discriminator used in this paper (the latter with the classification branches). Here, $d_{id}$, $d_{exp}$ and $d_{noise}$ are the dimensions for identity, expression and noise, respectively; $n_{id}$ is the number of distinct labels for identity, and $n_{exp}$ the number of distinct labels for expression. We use Leaky ReLU with a slope of $0.2$.

| Operation | Activation | Output Shape |
|---|---|---|
| $z \sim \mathcal{N}(0, I)$ | – | $d_{id} + d_{exp} + d_{noise}$ |
| Linear | LReLU | 512 |
| Linear | – | 66387 |
| Reshape | – | $22129 \times 3$ |

(a) Generator

| Operation | Activation | Output Shape |
|---|---|---|
| Input | – | $22129 \times 3$ |
| Geometry mapping | – | $3 \times 64 \times 64$ |
| **Common branch** | | |
| Conv $3 \times 3$ | LReLU | $16 \times 32 \times 32$ |
| Conv $3 \times 3$ | LReLU | $32 \times 16 \times 16$ |
| **Discriminator branch** | | |
| Conv $3 \times 3$ | LReLU | $64 \times 8 \times 8$ |
| Conv $3 \times 3$ | LReLU | $128 \times 4 \times 4$ |
| Reshape | – | 2048 |
| Linear | – | 1 |
| **Identity branch** | | |
| Conv $3 \times 3$ | LReLU | $64 \times 8 \times 8$ |
| Conv $3 \times 3$ | LReLU | $128 \times 4 \times 4$ |
| Reshape | – | 2048 |
| Linear | – | $n_{id}$ |
| **Expression branch** | | |
| Conv $3 \times 3$ | LReLU | $64 \times 8 \times 8$ |
| Conv $3 \times 3$ | LReLU | $128 \times 4 \times 4$ |
| Reshape | – | 2048 |
| Linear | – | $n_{exp}$ |

(b) Discriminator and Classifiers.

Figure 8: Generator and Discriminator used for experiments in the paper

## Decoupling Evaluation - Implementation Details

We train the embedding networks using a Resnet-18 architecture with input images of size $224 \times 224$. The images contain the orthographic projection of the facial mesh, and the values in the RGB channels encode the normal direction of each vertex, as we found this to give better results than the UV images. The networks were trained using the datasets described in Section 5.2 with the provided labels. The threshold is selected such that it maximizes the accuracy on the validation set, while keeping the False Acceptance Rate (FAR) below $10\%$. We build the validation set by randomly choosing an equal number of positive and negative pairs from the testing split. We choose $0.14$ as threshold for identity, which achieves $98.66\%$ accuracy and a FAR of $1.21\%$. For expression we use $0.226$ as threshold, which achieves $84.2\%$ of accuracy and a FAR of $8.03\%$.

## References


[1] Fanelli, Gabriele, et al. *"A 3-d audio-visual corpus of affective communication"*. IEEE Transactions on Multimedia 12.6 (2010): 591-598. 1

[2] McAuliffe, Michael, et al. *"Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi"*. Interspeech. 2017. 1

[3] Neti, Chalapathy, et al. *"Audio visual speech recognition"*. IDIAP. 2000. 1