EMOCA: Emotion Driven Monocular Face Capture and Animation **Supplementary Material**

Radek Daněček rdanecek@tue.mpg.de
Michael J. Black black@tue.mpg.de
Timo Bolkart tbolkart@tue.mpg.de
Max Planck Institute for Intelligent Systems, Tübingen, Germany

1. Introduction

Discussion of novelty: In Sup. Mat. we aim to shed more light on the process that eventually led to EMOCA and the challenges that had to be overcome. The idea of using deep perceptual losses to supervise face reconstruction is not new. A critic might argue, that the novelty of EMOCA is very limited for exactly that reason. However, the fact remains that previous SOTA methods have a clear limitation when it comes to reconstructing faces that communicate the correct emotional content. And from the knowledge of this limitation, we conceived the idea of leveraging emotion recognition, an idea not previously attempted by any work on face reconstruction. The inventive novelty in EMOCA was in coming up with the idea in the first place. This idea, once explained, makes such an intuitive sense, it may lead the reader into thinking it is a straightforward change to an already functioning system. The idea, although very simple and elegant, was by no means easy to get to work and this is what we aim to explain next.

Designing EMOCA: Our work starts with the simple idea - how can we employ the findings from emotion recognition to improve face reconstruction? Leveraging a pretrained SOTA network for emotion recognition, similarly to the way face recognition networks were used seems like a natural choice. However, using its final outputs such as the expression class and valence and arousal levels is not sufficient. Clearly, these very low-dimensional labels, while they do carry some information about the emotional content, they likely exhibit a lot of ambiguity and are not sufficient to supervise 3D shapes. For instance, an expression classified as happy can take on many different shapes (a subtle smile, a big smile with an open mouth, an "inverted" smile, etc.) and similar reasoning could be applied for any other expression and for any levels of valence and arousal as well. Hence, these labels most likely do not provide a sufficient supervision signal for geometry. The next logical design choice is to leverage high dimensional deep features from a pretrained emotion recognition network. This choice can only make sense if the emotion feature in question is a "well-behaved" embedding space. Ideally we want similar features to represent faces of similar expressions and vice versa. Therefore, we conducted an emotion retrieval experiment, using a pretrained publicly available EmoNet model [16] and nearest neighbors search. This experiment is discussed in Sec. 8. Having verified, that similar emotion features retrieve images of geometrically and semantically similar expressions, the next thing to be verified is whether the emotion feature carries a signal that is strong enough, to be utilizable for 3D reconstruction. This was particularly challenging and we comment on this further in Sec. 4. Finally, having demonstrated that the emotion recognition features indeed carry enough information in order to supervise the geometry, we can finally incorporate the emotion consistency loss into a face reconstruction framework, arriving at EMOCA. In addition to the ablations listed in the main paper, we also add ablations on different architectures and weights for the emotion consistency loss in Sec. 6

2. Implementation details

Emotion recognition metrics: In the main paper, we evaluate emotion metrics in the same setting as Toisoul et al. [16]. The metrics are defines as follows RMSE stands for root mean squared error:

$$\mathrm{RMSE}(Y, \hat{Y}) = \sqrt{\mathbb{E}[(Y - \hat{Y})^2]}.$$

SAGR stands for sign agreement and it evaluates whether the predicted value has the same sign as the ground truth:

$$\mathrm{SAGR}(Y, \hat{Y}) = \frac{1}{n} \sum_{i=1}^{n} \delta\left(\mathrm{sign}\left(y_i\right), \mathrm{sign}\left(\hat{y}_i\right)\right).$$

Pearson correlation coefficient (PCC) measures the correlation between predictions and GT:

$$\mathrm{PCC}(Y, \hat{Y}) = \frac{\mathbb{E}[(Y - \mu_Y)(\hat{Y} - \mu_{\hat{Y}})]}{\sigma_Y \sigma_{\hat{Y}}}.$$

Concordance correlation coefficient (CCC) incorporates the PCC but also penalizes signals which are still correlated

according to PCC but have different means:

$$\mathrm{CCC}(Y,\hat{Y}) = \frac{2\sigma_Y\sigma_{\hat{Y}}\,\mathrm{PCC}(Y,\hat{Y})}{\sigma_Y^2+\sigma_{\hat{Y}}^2+\left(\mu_Y-\mu_{\hat{Y}}\right)^2}.$$

Emotion recognition loss function: We train our emotion networks with the same loss function as defined by Toisoul et al. [16].

$$\mathcal{L}_{\text{categories}}\left(Y,\hat{Y}\right) = \text{ Cross entropy }\left(Y,\hat{Y}\right) = -\sum_{i=1}^{n}\hat{y}_i\log\left(y_i\right)$$

The complete loss function for emotion recognition is then defined as:

$$\mathcal{L}(Y,\hat{Y}) = \mathcal{L}_{\text{categories}}\left(Y,\hat{Y}\right) + \frac{\alpha}{\alpha+\beta+\gamma}\mathcal{L}_{\text{MSE}}(Y,\hat{Y}) + \frac{\beta}{\alpha+\beta+\gamma}\mathcal{L}_{\text{PCC}}\left(Y,\hat{Y}\right) + \frac{\gamma}{\alpha+\beta+\gamma}\mathcal{L}_{\text{CCC}}(Y,\hat{Y}),$$

where α, β and γ are shake-shake regularization coefficients [6] uniformly sampled from the interval $[0,1]$ for each training batch and:

$$\mathcal{L}_{\text{MSE}}(Y,\hat{Y}) = \text{MSE}_{\text{valence}}\left(Y,\hat{Y}\right) + \text{MSE}_{\text{arousal}}\left(Y,\hat{Y}\right)$$

$$\mathcal{L}_{\text{PCC}}(Y,\hat{Y}) = 1 - \frac{\text{PCC}_{\text{valence}}\left(Y,\hat{Y}\right) + \text{PCC}_{\text{arousal}}\left(Y,\hat{Y}\right)}{2}$$

$$\mathcal{L}_{\text{CCC}}(Y,\hat{Y}) = 1 - \frac{\text{CCC}_{\text{valence}}\left(Y,\hat{Y}\right) + \text{CCC}_{\text{arousal}}\left(Y,\hat{Y}\right)}{2}.$$

Unlike the work of Toisoul et al. [16], we do not use knowledge distillation as its improvements are marginal and make the training process much more complex.

Image-based emotion recognition: We investigate emotion recognition networks based on different architectures, ResNet-50 [8], Swin Transformer [12], and EmoNet [16]. We train all models on AffectNet [13], using the training/validation/test split proposed by Toisoul et al. [16]. The ResNet-50 and Swin Transformer based models are pretrained on ImageNet [3]. During training, the training images are sampled such that each of the 7 expression labels appears with the same frequency. This sampling is crucial to maximize the performance of the emotion networks, as the AffectNet training set is not balanced. We use the Adam optimizer with learning rate of 0.0001, β_1 = 0.9 and β_2 = 0.999. The batch size used for training is 64. Each model is trained for a maximum of 20 epochs with early stopping, and the model with the lowest validation error is selected.

3DMM-based emotion recognition:

In Section 5.2 of the paper (Tab. 1) and Table 1, we evaluate different face reconstruction methods by recognizing emotions from the regressed 3DMM parameters. Specifically, we train a 4-layer MLP with Batch Normalization and LeakyReLUs to output valence and arousal levels and expression classes from the regressed identity and expression parameters (see Figs. 1 and 2 for details). The size of each hidden layer is 2048. We train the 3DMM-based recognition on AffectNet similarly to the image-based emotion recognition. The loss function is identical to the one used for image-based emotion recognition. The batch size used for training is 64. We use the Adam optimizer with a learning rate of 0.0001, $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

Detail stage training: The detail stage training follows the training protocol of DECA [5]. The coarse model part is kept fixed, while detail encoder and decoder are trained. This stage uses VGGFace2 [1] and VoxCeleb2 [2] images, due to the necessity of having multiple images per identity. We optimize following losses: photometric loss, ID-MRF perceptual loss which encourages reconstruction of higher frequency detail (compared to the coarse mesh), as well as the soft symmetry loss and displacement regularization. Further, to disentangle identity and expression dependent details, we employ DECA's detail consistency loss, where each batch contains k images of each subject, and the detail codes are exchanged randomly between the predictions for each identity For our training, we set k=3 and batch size of 4 identities, totalling 12 input images per batch. For more details, see the original DECA publication.

3. Qualitative evaluation

In addition to the performance in emotion analysis on the AffectNet dataset in the main paper, we also test EMOCA on AFEW-VA [10]. The results are reported in Tab. 1.

4. Emotion optimization

We can use our emotion consistency loss for additional tasks. Here we consider the problem of expression retargeting. Given two face images, a source identity image I_S and a target expression image I_T of potentially two different people with different expressions, poses, cameras, and lighting, our goal is to optimize for the (unknown) target expression $\hat{\boldsymbol{\psi}}_T$. Formally, we infer the FLAME parameters $E_c(I_S)$ and $E_c(I_T)$ for both images. Then, with some abuse of notation, we render $I_R(\boldsymbol{\psi}) = R(M(\boldsymbol{\beta}_S, \boldsymbol{\theta}_T, \boldsymbol{\psi}), \boldsymbol{\alpha}_S, \mathbf{l}_T, \mathbf{c}_T)$, the FLAME mesh with source identity shape $\boldsymbol{\beta}_S$, source albedo $\boldsymbol{\alpha}_S$, and target pose $\boldsymbol{\theta}_T$, target camera $\mathbf{c}_T$, target lighting $\mathbf{l}_T$, and the optimization expression parameters $\boldsymbol{\psi}$. We then extract the emotion features of the rendering $\boldsymbol{\epsilon}_R(\boldsymbol{\psi}) = A(I_R(\boldsymbol{\psi}))$ and the target image $\boldsymbol{\epsilon}_T = A(I_T)$, and optimize:

$$\hat{\boldsymbol{\psi}}_T = \arg\min_{\boldsymbol{\psi}} d(\boldsymbol{\epsilon}_R(\boldsymbol{\psi}), \boldsymbol{\epsilon}_T) + \lambda_{\boldsymbol{\psi}} L_{\boldsymbol{\psi}}, \tag{1}$$

with $d(\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2) = \|\boldsymbol{\epsilon}_1 - \boldsymbol{\epsilon}_2\|_2$, expression regularizer $L_{\boldsymbol{\psi}} = \|\boldsymbol{\psi}\|_2^2$, and regularizer weight $\lambda_{\boldsymbol{\psi}} = 1e{-}3$. We use gradient descent for the optimization. Below we show optimization

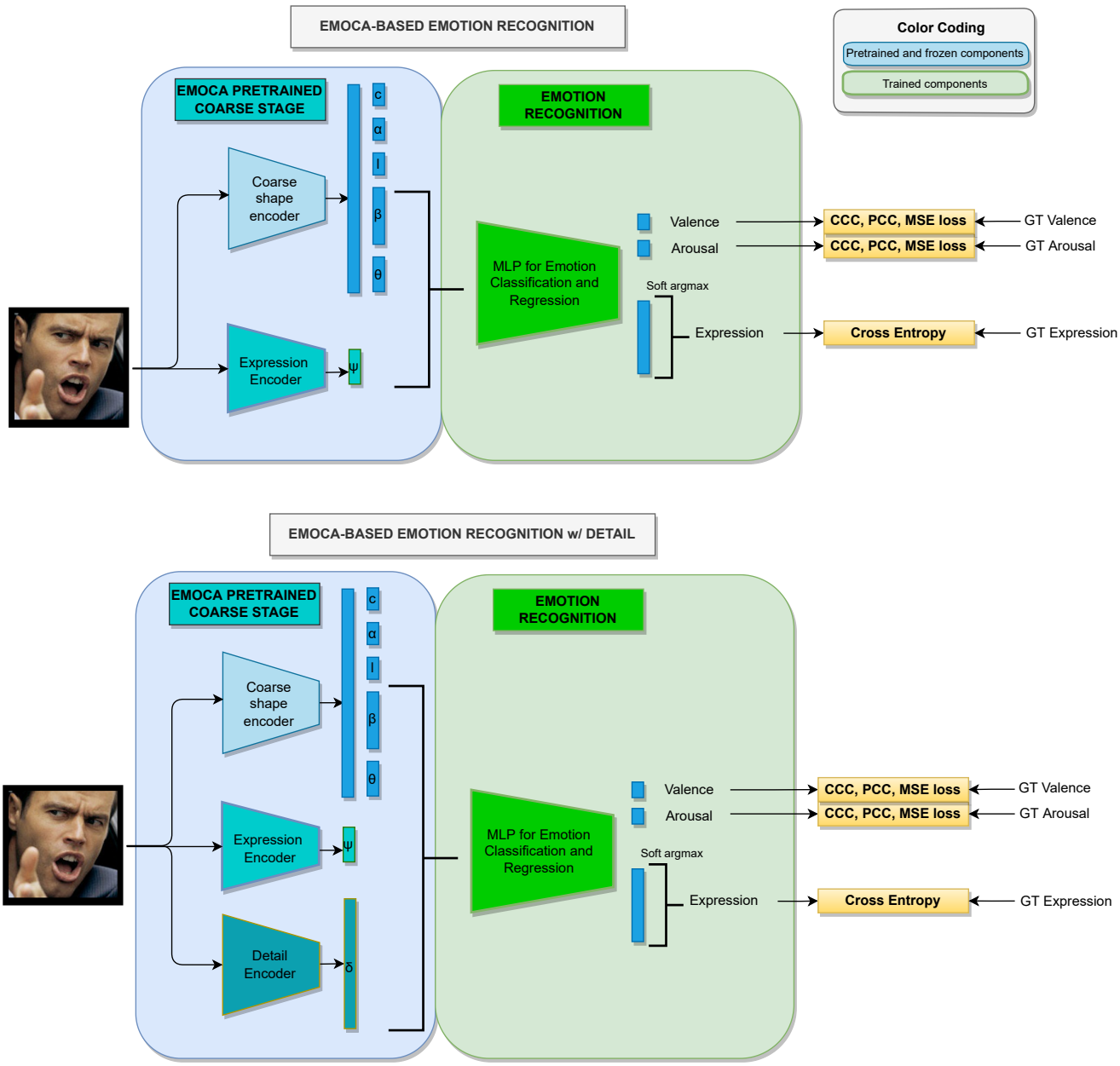

Figure 1. The architecture of EMOCA-based emotion recognition. Top: EMOCA emotion recognition with coarse parameters. From the pretrained coarse stage we extract the shape parameters $\boldsymbol{\beta}$, expression parameters $\boldsymbol{\psi}$ and jaw pose $\boldsymbol{\theta}_{jaw}$. A similar approach is taken for DECA-based recognition, except that DECA does not have a dedicated expression encoder. These are fed to an MLP to regress valence and arousal and classify expression. Bottom: emotion recognition for EMOCA-based reconstruction methods with detail code included.

results, and an analysis of the convergence and sensitivity to the initialization.

On Emotion Network Architecture: Figure 3 shows emotion optimization results using different emotion recognition network. This indicates that the original released EmoNet is not suitable for emotion optimization. Instead, we use the ResNet-50 architecture as default model.

On Initialization: Figure 4 further shows the influence of the initialization on the optimized emotion. These results demonstrate that 3DMMs, when rendered, can in fact be animated with a deep perceptual emotion similarity loss.

On Jaw Optimization: A perceptive reader may ask, why we optimize only for the expression parameters $\boldsymbol{\psi}$ and not also for the jaw pose $\boldsymbol{\theta}_{jaw}$. After all, the jaw position most certainly has an effect on the perceived emotion. We have struggled with the jaw optimization issue for quite a long

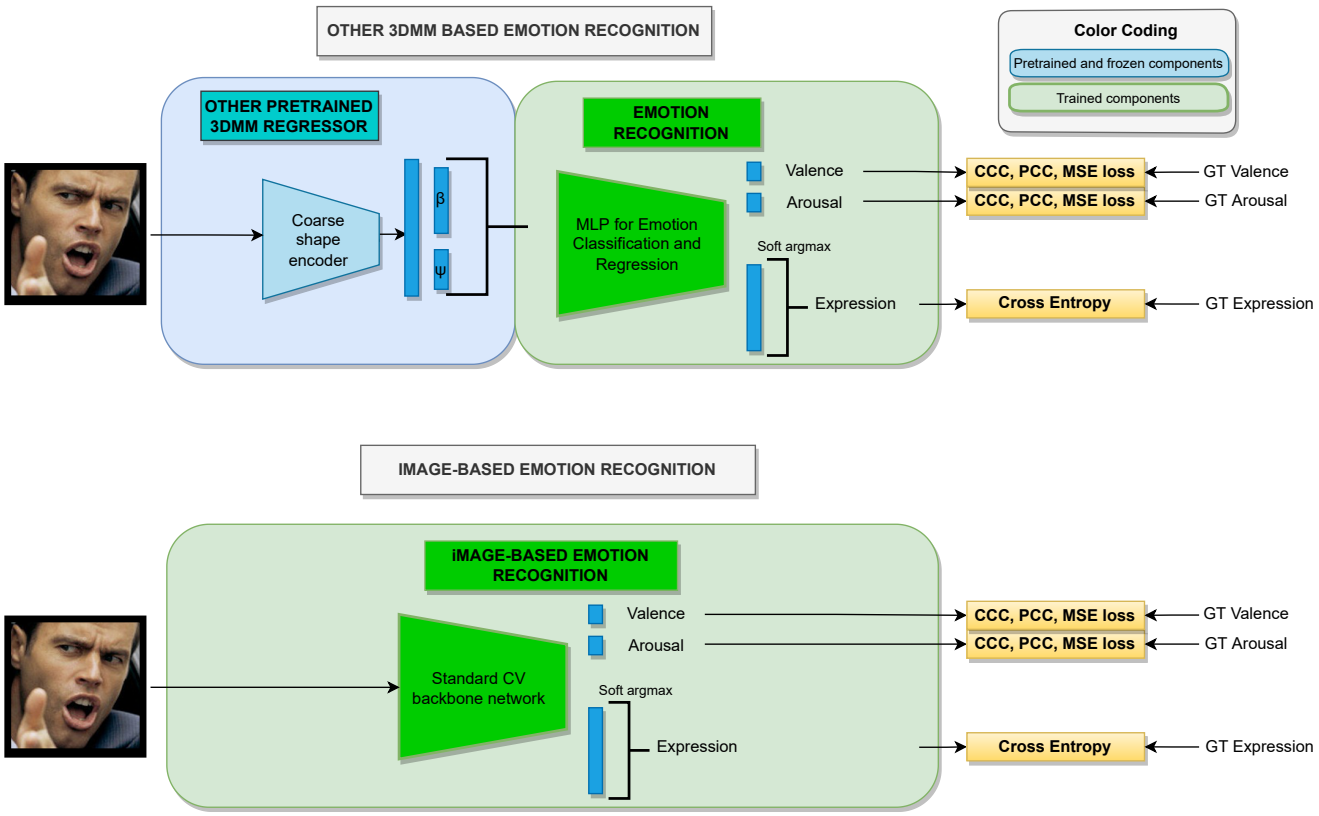

Figure 2. The architecture of other emotion recognition netowrks. Top: emotion recognition for other 3DMM-based reconstruction methods (Deep3DFace [4], 3DDFA-V2 [7], MGCNet [15]. These have a single decoder that regress to the Basel Face Model [14] parameter space, which does not model jaw pose explicitly. Therefore only β and ψ are considered. Bottom: a standard image-based network trained for emotion recognition. Both types of emotion recognition are trained with the same supervision.

Model	V-PCC ↑	V-CCC ↑	V-RMSE ↓	V-SAGR ↑	A-PCC ↑	A-CCC ↑	A-RMSE ↓	A-SAGR ↑
EmoNet	0.59	0.54	0.22	0.61	0.55	0.49	*0.22*	*0.80*
Deep3DFace	*0.64*	*0.59*	*0.21*	**0.65**	0.55	0.48	**0.21**	**0.81**
ExpNet	0.31	0.25	0.27	0.55	0.36	0.30	0.24	0.79
MGCNet	0.54	0.50	0.23	0.62	0.49	0.44	0.23	0.79
3DDFA	0.41	0.38	0.27	0.57	0.44	0.41	0.24	0.78
DECA (coarse)	0.57	0.53	0.23	0.62	*0.55*	*0.50*	*0.22*	**0.81**
DECA /w detail	0.57	0.53	0.23	0.63	0.53	0.49	*0.22*	*0.80*
EMOCA (Ours)	**0.65**	**0.63**	*0.21*	*0.64*	**0.57**	**0.54**	*0.22*	*0.80*
EMOCA /w detail (Ours)	**0.68**	**0.65**	**0.20**	*0.64*	*0.56*	*0.53*	*0.22*	*0.80*

Table 1. Emotion recognition performance on AFEW-VA [10]. All emotion regressors are pretrained on AffectNet and finetuned on the AFEW-VA using 5-fold Cross-Validation (CV). The reported numbers are averaged across the 5-fold CV runs. EMOCA performs best, followed by Deep3DFace. Surprisingly, both of these methods outperform EmoNet. Other 3D-based methods follow.

time, unable to get acceptable results as the jaw pose parameter optimization makes this optimization unstable - the jaw would always be posed to an unrealistic or at least very incorrect pose. Fixing the jaw pose to a reasonable estimate however (such as DECA's prediction) makes the optimization stable and produces good results. We hypothesise that this instability could be caused by the following:

1. FLAME is missing a comprehensive prior for the jaw pose. We experimented with simplistic hand-crafted priors (such as distance or squared distance from the expected pose) but this did not yield any improvement. It is possible that the creating a more comprehensive prior (other than the Gaussian prior for FLAME's expression space), a prior that entangles the expression and jaw pose spaces is necessary. This makes for an interesting direction for future work.

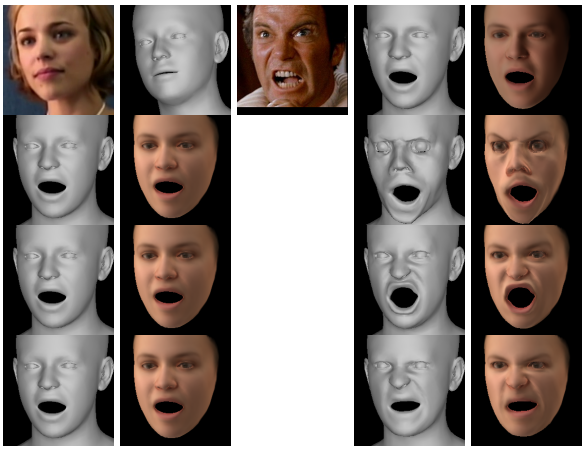

Figure 3. Emotion optimization example. The first row contains a source image, its DECA reconstruction, a target image, its DECA reconstruction, and the colored reconstruction. The following rows contain: the initialization of the optimization w/o and w/ color (left) and the optimization result w/o and w/ color (right). The different rows use different emotion recognition networks for optimization. The second row uses the original released EmoNet, the third row a self-trained EmoNet, and the bottom row using our ResNet-50 model. While EmoNet gives SOTA emotion recognition results, it is less suitable for our task of emotion-driven expression optimization or reconstruction.

2. Emotion optimization involves optimizing a deep feature vector and while we have demonstrated that similar emotion features belong to similar expressions, we have not eliminated the possibility, that the emotion network can be "attacked" to produce the desired features with a distorted images. An optimization process, in which the jaw is not fixed could results in an adversarial attack on the network that forces it to produce a similar emotion feature vector.

5. Perceptual study

Section 5.2 of the paper evaluates the amount emotion conveyed by the reconstructed 3D geometry in a perceptual study. Figure 5 gives the full confusion matrix of the participants' labels of real images (rows) and the labels of the reconstructions (columns). Figure 6 further compares the ground truth emotion labels with the participants' classifications of the reconstructions. For completeness, we also include the confusion matrix of participants' labeling of the real images in Fig. 7.

6. Emotion consistency

Emotion network architecture: The choice of architecture for emotion supervision plays a critical role. While all architectures perform comparatively well on the emotion recognition task, they are not equally suitable as supervision for our 3D face reconstruction task. Fig. 8 visually compares EMOCA models trained with different emotion recognition networks as supervision. Again, the SOTA emotion recognition architecture - EmoNet, is not suitable as it produces unacceptable artifacts. Furthermore, the SWIN [12] transformer backbone, which is considered to be superior to the ResNet [8] architecture, also produces some undesirable artifacts. Hence, the ResNet backbone was used for the final model of the emotion recognition network.

Figure 4. Sensitivity of the emotion optimization to initialization. The first row contains a source image, its DECA reconstruction, a target image, its DECA reconstruction, and the colored reconstruction. The following rows contain: the initialization of the optimization w/o and w/ color (left) and the optimization result w/o and w/ color (right). Note that the optimization process is only modifying the expression coefficients $\boldsymbol{\psi}$ and not the jaw rotation $\boldsymbol{\theta}_{jaw}$. While the process usually converges to meaningful results, the most favorable outcome is obtained, when initializing the process with the target expression coefficients $\boldsymbol{\psi}$ and pose $\boldsymbol{\theta}$, which correspond to the second row.

Emotion consistency weight: We have experimented with different values of the emotion consistency loss weight term λ_{emo}. This is a crucial factor of successfully train-

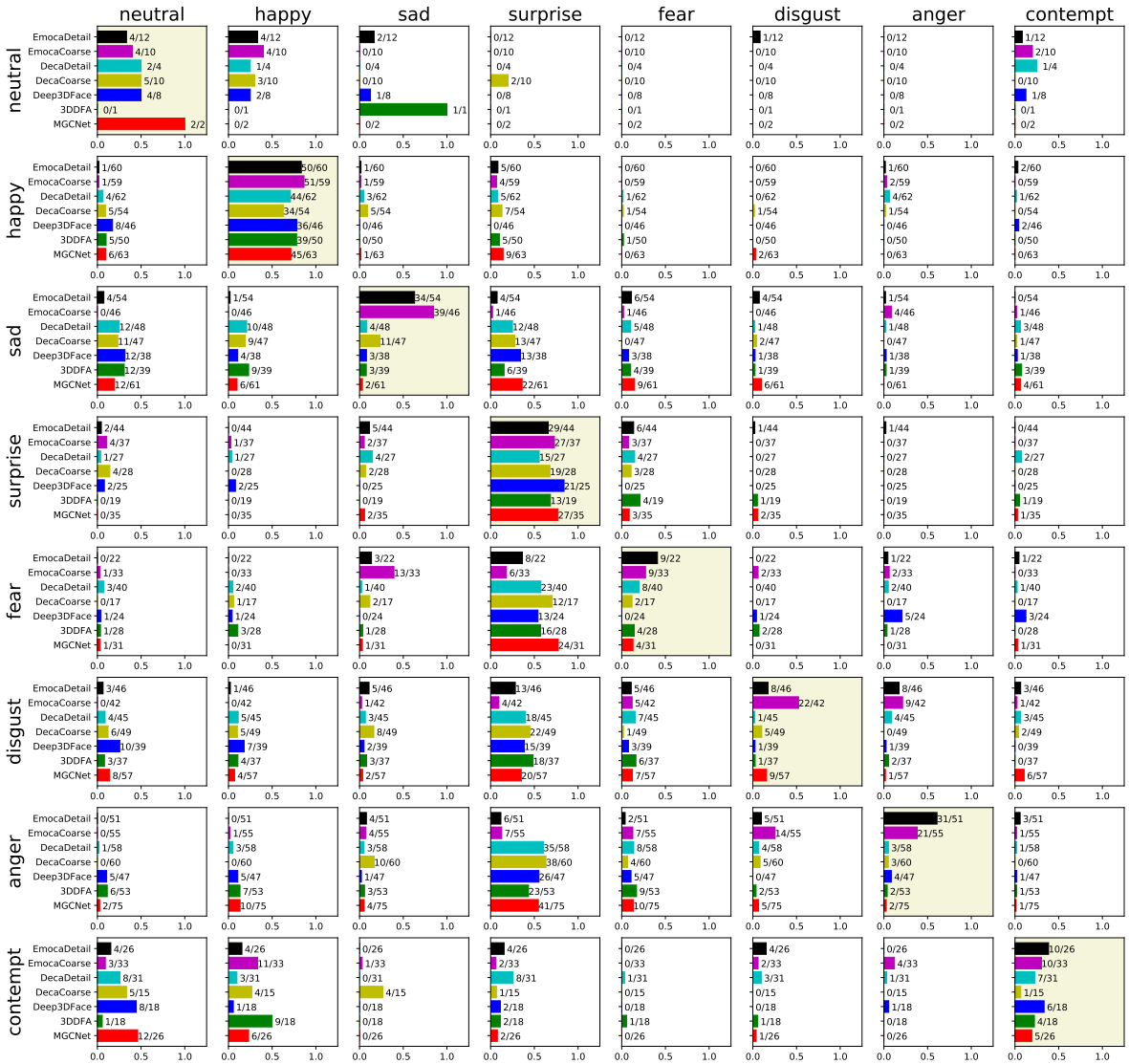

Figure 5. This figure contains the confusion matrices of participant's labels of the real image and the reconstructed images for each method. The x-axis of each cell gives the ratio of participants' reconstruction labels and real image labels and the absolute number is written next to each bar. The accuracy of each method for a particular expression class is on the diagonal. You can see that both variants of EMOCA (detail and coarse) are superior to the other methods. Furthermore, off-diagonal you can observe how the label of meshes reconstructed by EMOCA is much less confused for other labels, compared to other methods. Finally, the confusion matrix highlights how other methods are not capable of producing expressions of fear, disgust and anger. Instead these are confused with surprise. EMOCA does not suffer from the same limitation. However, participants did have some trouble distinguishing reconstructions of disgust and anger. Please note that the first row (neutral) shows a small number of samples. This happens because our perceptual study did not contain neutral images.

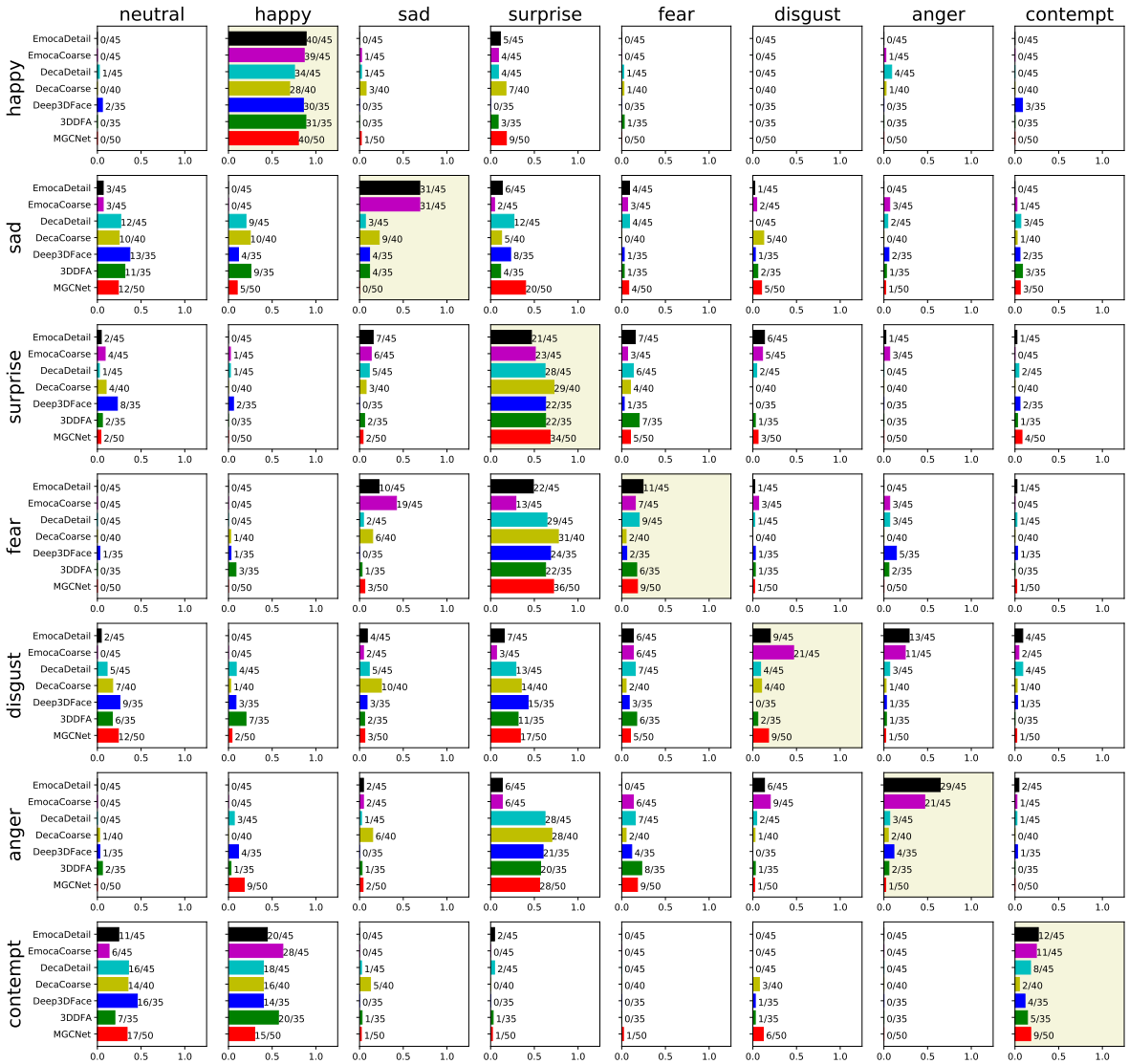

Figure 6. This figure contains the confusion matrices of participant's labels of the reconstructions w.r.t. to the ground truth labels (as opposed to users' subjective labels, which you can find in Fig. 5. Please note that neutral expressions were not given in the study, which is why the matrix only has six rows (neutral excluded).

ing EMOCA. If the weight is too small, the emotion is not captured well enough. At the same time, high values lead to unnaturally over-exaggerated expressions. A visual ablation of this phenomenon can be found in Fig. 9 and Fig. 10 for two different emotion network architectures; ResNet-50 [8] and SWIN-B [12].

Additional ablations: We further evaluate the impact of the similarity metric used for the emotion similarity, the effect

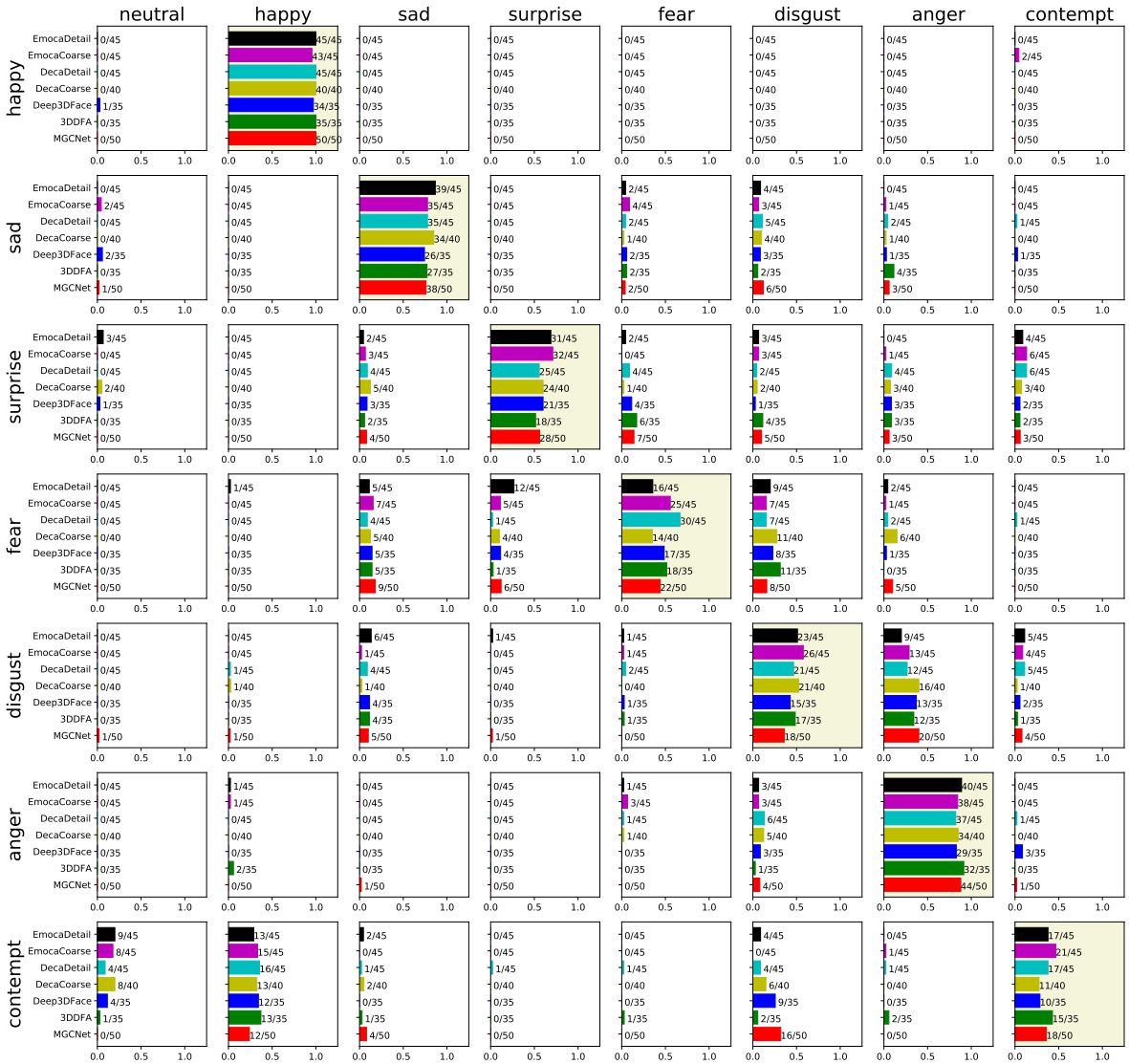

Figure 7. This figure contains the confusion matrices of participant's labels of the real images w.r.t. to ground truth images. While this figure does not compare the performance of methods, it serves as a baseline comparison to Fig. 6. Classifying expression is subjective. While our participants mostly agreed with our ground truth, there were disagreements for the negatively charged expressions of fear, disgust, anger and particularly contempt.

of adding a landmark reprojection error to the loss function, and the effect of the relative landmark losses (mouth closure, lip corner distance and eye closure). Finally, we analyze the effect of using DECA's training data instead of AffectNet. You can see the results in Fig. 11.

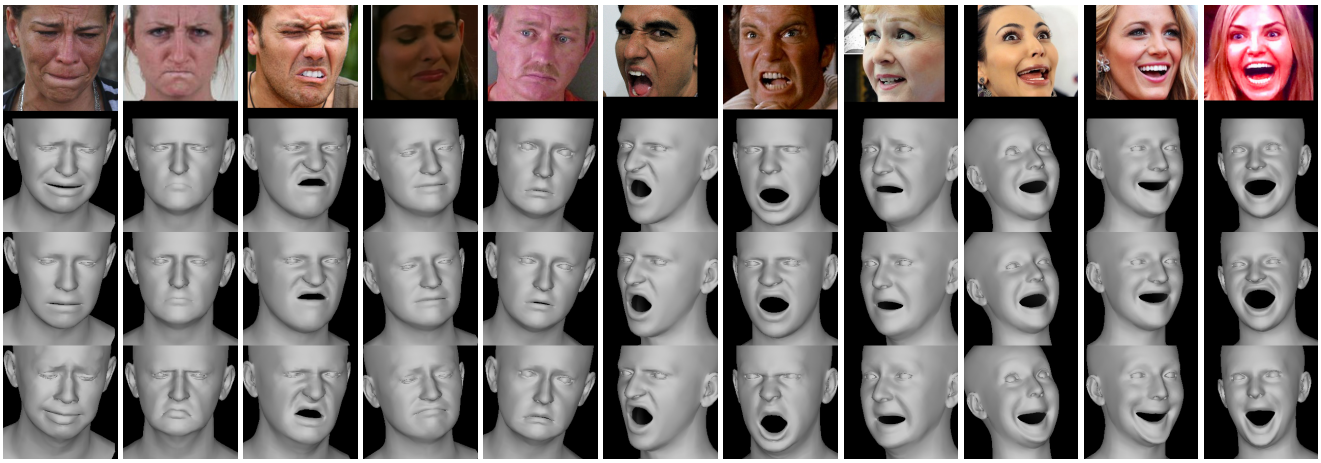

Figure 8. Comparison of different EMOCA models, supervised by different emotion networks. From top to bottom: ResNet-50 [8], SWIN-B [12], EmoNet [16]. All three networks affect the reconstruction in different ways. EMOCA-ResNet produces the best visual results and is our model of choice. EMOCA-SWIN produces results of slightly lower visual quality. Finally, EMOCA-EmoNet sometimes produces unrealistic expressions, which makes EmoNet less suitable for this task.

Figure 9. Comparison of models trained with different weights of the emotion consistency loss λ_{emo}. The emotion network used was ResNet-50 [8]. Top row consists of input images. Different values of λ_{emo} follow. From top to bottom 0, 0.1, 0.5, 1 (final EMOCA), 5, 10.

7. Emotional retargeting

EMOCA regresses FLAME [11] parameters and expression dependent geometric details. The disentanglement of the coarse identity and expression geometry and the identity and expression dependent details allows us to animate EMOCA's reconstructions. We demonstrate this by animating a source 3D face using a video sequence of another actor. Figure 12 demonstrates two things, first, EMOCA re-

Figure 10. Comparison of models trained with different weights of the emotion consistency loss λ_{emo}. The emotion network used was SWIN-B [12]. Top row consists of input images. Different values of λ_{emo} follow. From top to bottom 0, 0.1, 0.5, 1 , 5, 10. While SWIN-B suffers from fewer artifacts compared to ResNet-50 when changing the weight, we have deemed the visual quality of results produce by a ResNet-supervised EMOCA slightly better, which is why ResNet was selected for the final model.

constructions convey emotions of the source images, and second, the animated faces of other subjects convey the same emotion. The emotional fidelity hence is preserved in the animated face of the other subject.

8. Emotion retrieval

Our work relies on the following key hypothesis. The emotion recognition networks learn a useful embedding of emotion. The following properties are desirable:

- Images of faces with similar expressions conveying similar emotions are close in this embedding space.
- Images of faces with dissimilar expressions/emotions are farther apart in this space.
- Invariance to pose, identity and lighting and background.

We employ the publicly released model of EmoNet [16] and use the 256-dimensional feature output of the last convolutional layer as emotion embedding. We then extract the emotion embedding for faces in the Aff-Wild2 video dataset [9]. For the emotion retrieval given an image, we seek the nearest neighbors w.r.t. L2 distance metric in the dataset. Figure 13 shows the 10 nearest neighbors for multiple images. For comparison, we repeat the process for the ground truth (GT) valence and arousal labels of the Aff-Wild2 dataset in Fig. 14.

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 67–74, 2018. 2

[2] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep speaker recognition. In *Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1086–1090, 2018. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 2

[4] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with

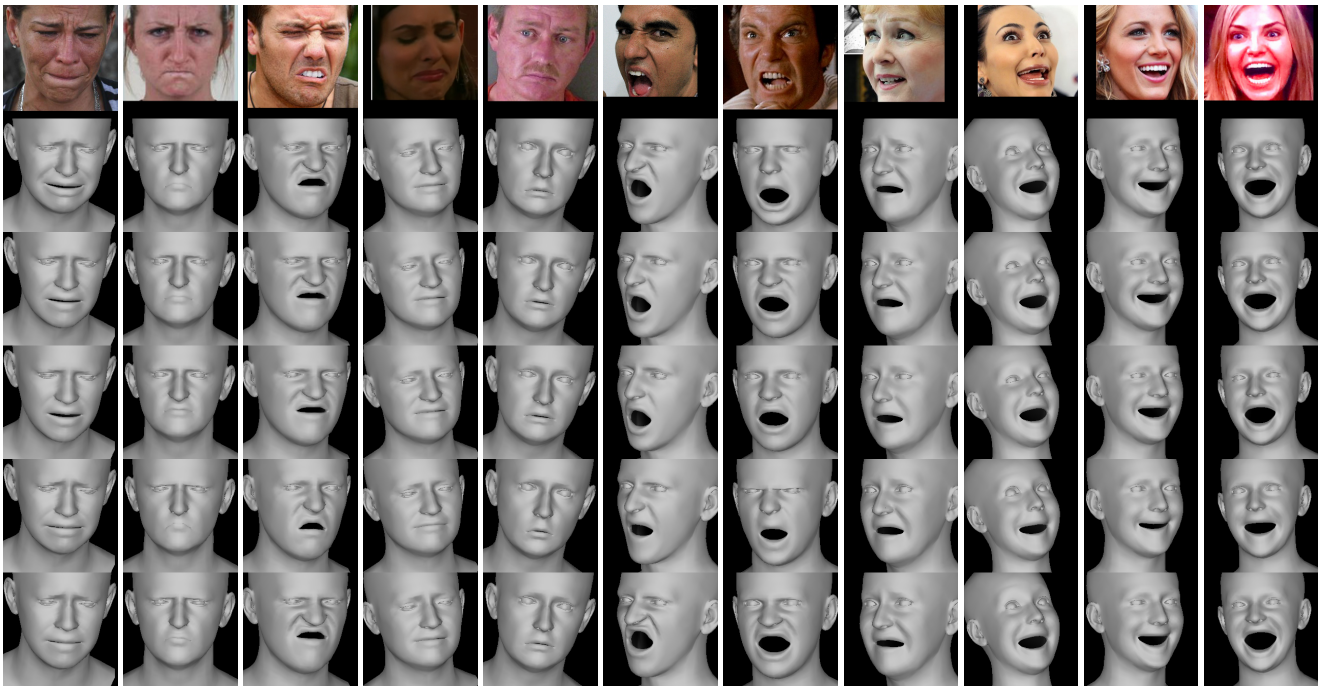

Figure 11. A visual comparison of model with different changes. First row consists of input images. The next three rows use different metrics for evaluating emotion similarity - L2 (EMOCA), L1 and cosine similarity. As you can observe, the selection of the metric is not critical for performance. The following row drops the relative landmark losses (mouth closure, eye closure and lip corner distance). Observe that this has a negative effect on the samples, particularly the mouth region. Final row is EMOCA model trained on the same data as DECA instead of AffectNet. You can see that it achieves a very similar result compared to EMOCA trained on AffectNet. This highlights an interesting finding - once an emotion recognition network has been trained, it can be used for supervision even on datasets that do not strictly guarantee a balanced representation of emotional states, such as face recognition datasets.

weakly-supervised learning: From single image to image set. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, pages 285–295, 2019. 4

[5] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *Transactions on Graphics, (Proc. SIGGRAPH)*, 40(4):88:1–88:13, 2021. 2

[6] Xavier Gastaldi. Shake-shake regularization. *CoRR*, abs/1705.07485, 2017. 2

[7] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3D dense face alignment. In *European Conference on Computer Vision (ECCV)*, pages 152–168, 2020. 4

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2, 5, 7, 9

[9] Dimitrios Kollias and Stefanos Zafeiriou. Aff-Wild2: Extending the Aff-Wild database for affect recognition. *CoRR*, abs/1811.07770, 2018. 10, 15

[10] Jean Kossaifi, Georgios Tzimiropoulos, Sinisa Todorovic, and Maja Pantic. Afew-va database for valence and arousal estimation in-the-wild. *Image and Vision Computing*, 65:23–36, 2017. 2, 4

[11] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 9

[12] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021. 2, 5, 7, 9, 10

[13] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 2

[14] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *International Conference on Advanced Video and Signal based Surveillance (AAAI)*, pages 296–301, 2009. 4

[15] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. In *European Conference on Computer Vision (ECCV)*, volume 12360, pages 53–70, 2020. 4

[16] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous va-

lence and arousal levels from faces in naturalistic conditions. 14
Nature Machine Intelligence, 3(1):42–50, 2021. 1, 2, 9, 10,

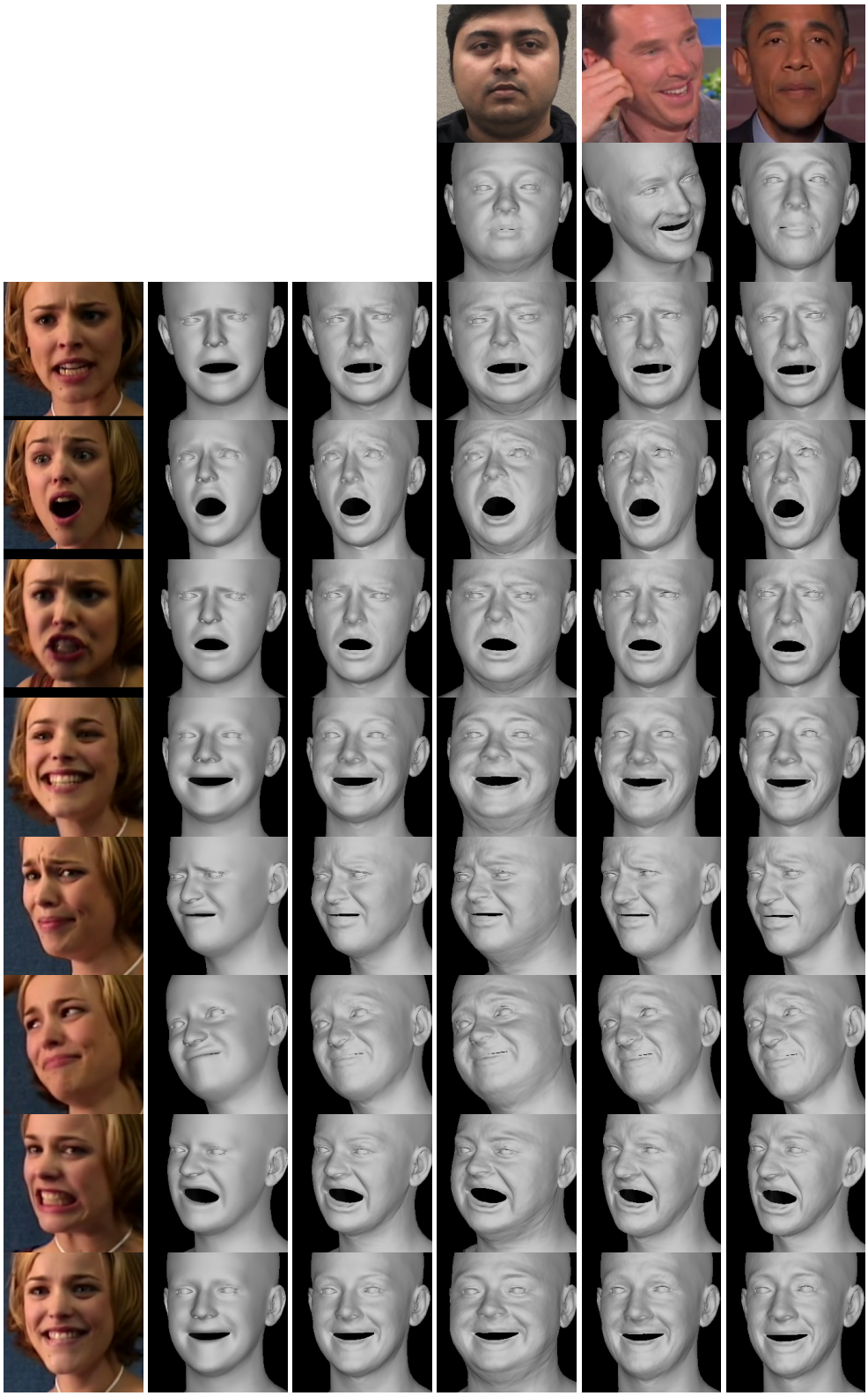

Figure 12. **Emotional retargetting.** From left to right. The input image, coarse reconstruction, detailed reconstruction, emotion retargeted to the coarse identity above. Observe that while the identity and the person-specific detailed displacements change with the source actor, the emotion fidelity is preserved. For the entire sequence in motion, please see the supplementary video.

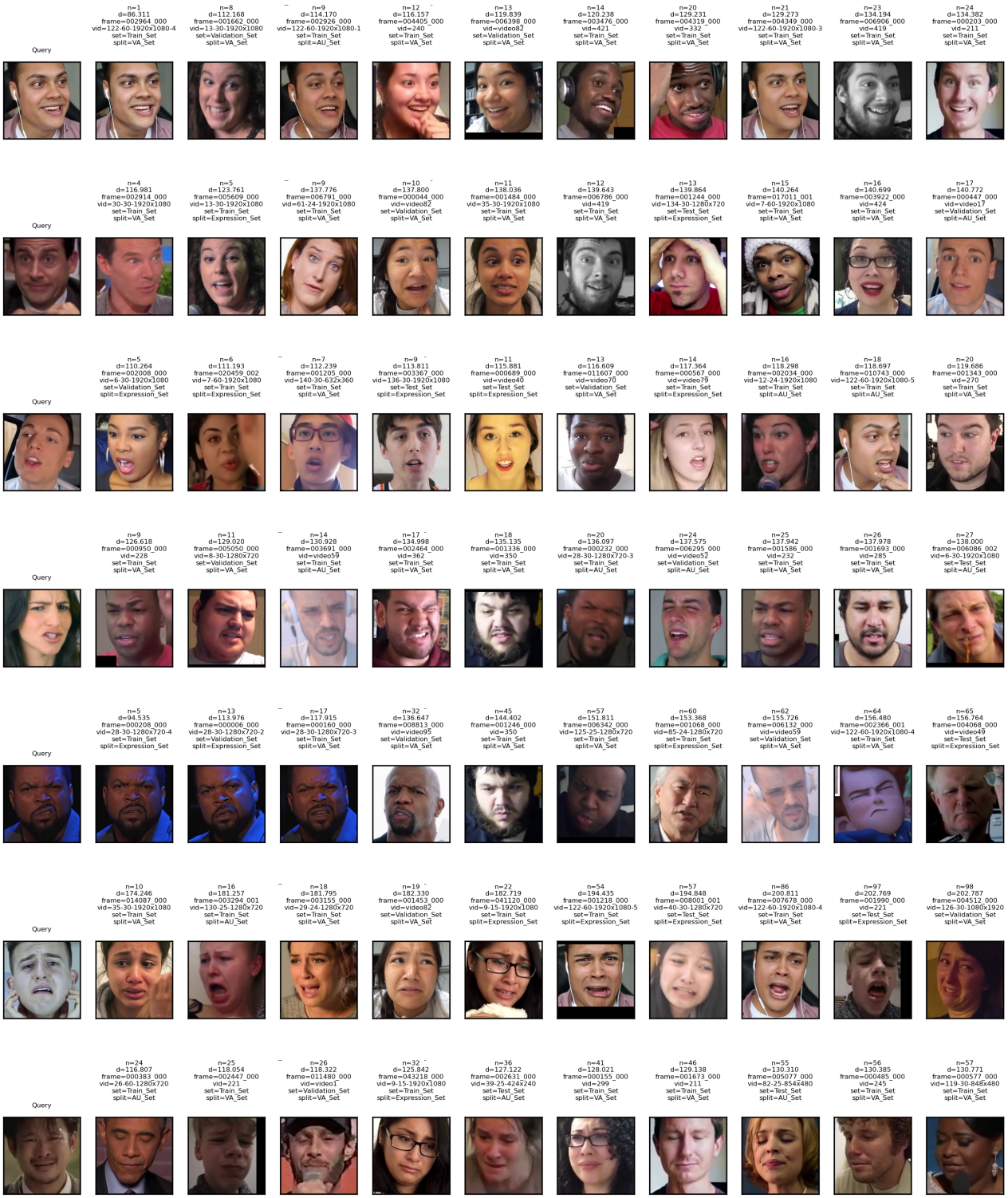

Figure 13. Examples of nearest neighbor retrieval using the EmoNet [16] feature. We searched for up to 100 neighbors. We ony include up to 1 NN per video to avoid retrieving consecutive frames. Left: query image, Right: ordered nearest neighbors from different clips. Observe how all of the retrieved faces communicate very similar emotional content.

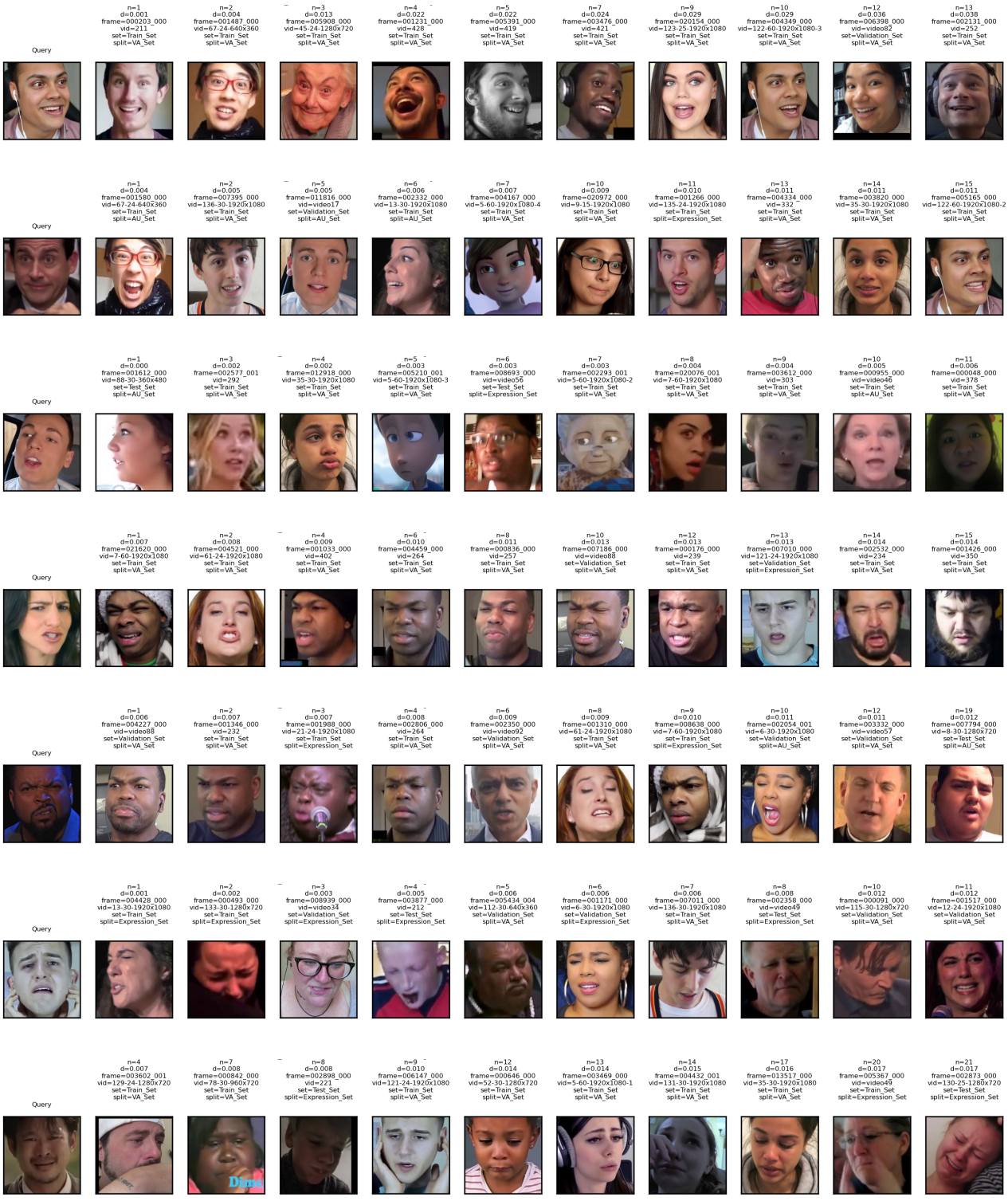

Figure 14. Examples of nearest neighbor retrieval using the ground truth annotated valence and arousal space on the AffWild2 [9] dataset. While the retrieved faces do have some degree of similarity, the quality of retrieval compared to the EmoNet feature is lower.