Supplementary Material - Attention Driven Cropping for Very High Resolution Facial Landmark Detection

Prashanth Chandran[1,2], Derek Bradley[2], Markus Gross[1,2], and Thabo Beeler[2]

[1]Department of Computer Science, ETH Zurich
[2]DisneyResearch|Studios, Zurich
chandrap@inf.ethz.ch, derek.bradley@disneyresearch.com, grossm@inf.ethz.ch, thabo.beeler@gmail.com

1. Overview

We present additional quantitative and qualitative results to show the benefits of our method against existing state of the art.

2. Additional Quantitative Results

2.1. Improvement across Image Resolutions

With high resolution imagery becoming more and more common, there will be an increasing need for algorithms which can process images of high quality. In Fig. 1, we compare the improvement in overall AUC by switching to our method when compared to DLIB [4], and FAN [1] as the resolution of the input image increases. At lower resolutions, the improvement though present is negligible. However, as the resolution of the input increases to 1K and beyond, the benefits of switching to our method substantially increase.

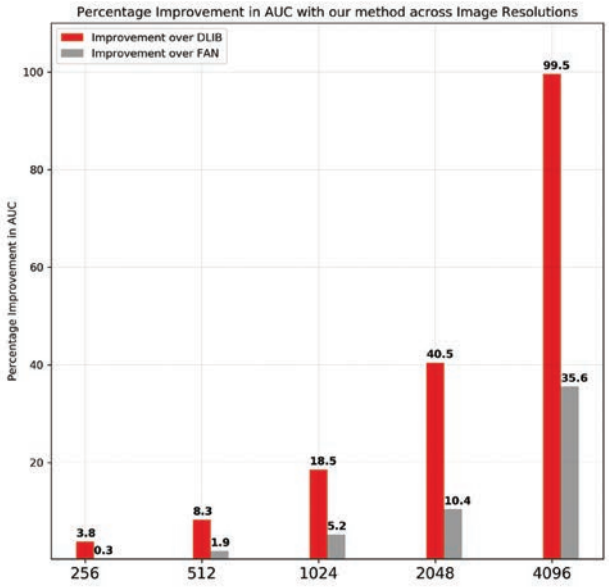

Figure 1. The benefit of our method increases with the resolution of the input image. At resolutions as low as 256 x 256, our method remains comparable to FAN [1]. However, a higher resolutions, we significantly outperform existing methods.

2.2. Regional Metrics

Our region based refinement results in an improvement in the quality of the predicted landmark for every considered region. In Fig. 2, we present the PCK and NME metric measured region-wise for resolutions ranging from 256 x 256 pixels to 2048 x 2048 pixels. As seen in the figure, our method improves results in every region across all resolutions.

3. Additional Qualitative Results

3.1. Visualizing regional refinement

A primary difference of the proposed method against existing multi-stage convolutional architectures are the regional refinement modules. Conventionally, multi-stage architectures [1, 6, 5, 2] feed the predictions from the previous stage as an additional input to the current stage, while also holistically considering the entire face as one region. While such stacking can also be performed with our architecture (see the ablation study in section 4.4 in the main paper), we only pass regional crops of the high resolution image to the regional hourglass modules. The predictions made by these regional hourglass are of much higher precision than those made by our global model. In Fig. 3 and Fig. 4, we visualize the predictions of our global and regional models on different expressions of a test subject.

3.2. Performance across viewpoints

It is well known that the performance of facial landmark detectors is view dependent. As existing databases consist of more images of people in frontal views as compared to profile views, methods such as [3] perform pose-based data balancing to obtain consistent performance across many different viewpoints. Since our training data was already pose balanced, we did not have to explicitly perform such data pre-processing. In Fig. 5 and Fig. 6, we compare the performance of DLIB [4], a 4 stage hourglass [1] and our method for two expressions from 4 different viewpoints. Our method obtains the lowest MSE for both expressions in all 4 viewpoints.

3.3. Performance across expressions

Finally, in Fig. 7, Fig. 8, Fig. 9, Fig. 10, and Fig. 11, we show qualitative comparisons of DLIB [4], the 4 stage hourglass [1] and our method for many different expressions of a test subject. Our method results in the least MSE for all expressions.

References

[1] A. Bulat and G. Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230, 000 3d facial landmarks). *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1021–1030, 2017. 1, 2, 6

[2] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CoRR*, abs/1611.08050, 2016. 1

[3] Z. Feng, J. Kittler, M. Awais, P. Huber, and X. Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. *CoRR*, abs/1711.06753, 2017. 2

[4] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014. 1, 2, 6

[5] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, pages 483–499, 2016. 1

[6] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4724–4732, 2016. 1

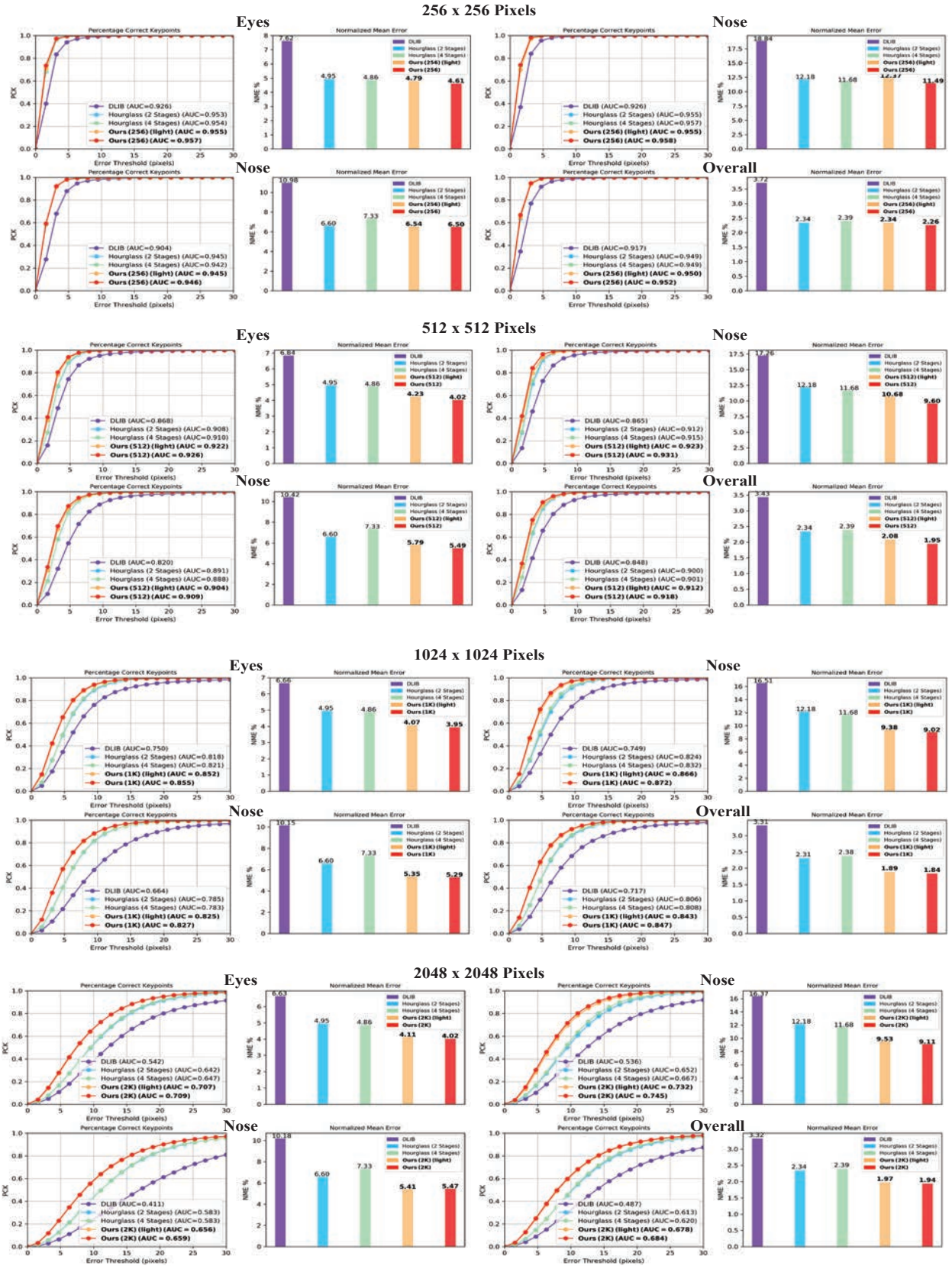

Figure 2. Percentage Correct Keypoints (PCK) and Normalized Mean Error (NME) are measured individually for different regions of the face for resolutions ranging from 256 x 256 to 2048 x 2048 (Resolutions increase a factor of 2 from top (256 x 256 pixels) to bottom (2048 x 2048 pixels)). Inside each row, metrics for individual regions of the face are presented (Eyes - top left, Nose - top right, Mouth - bottom left and Overall - bottom right)

.

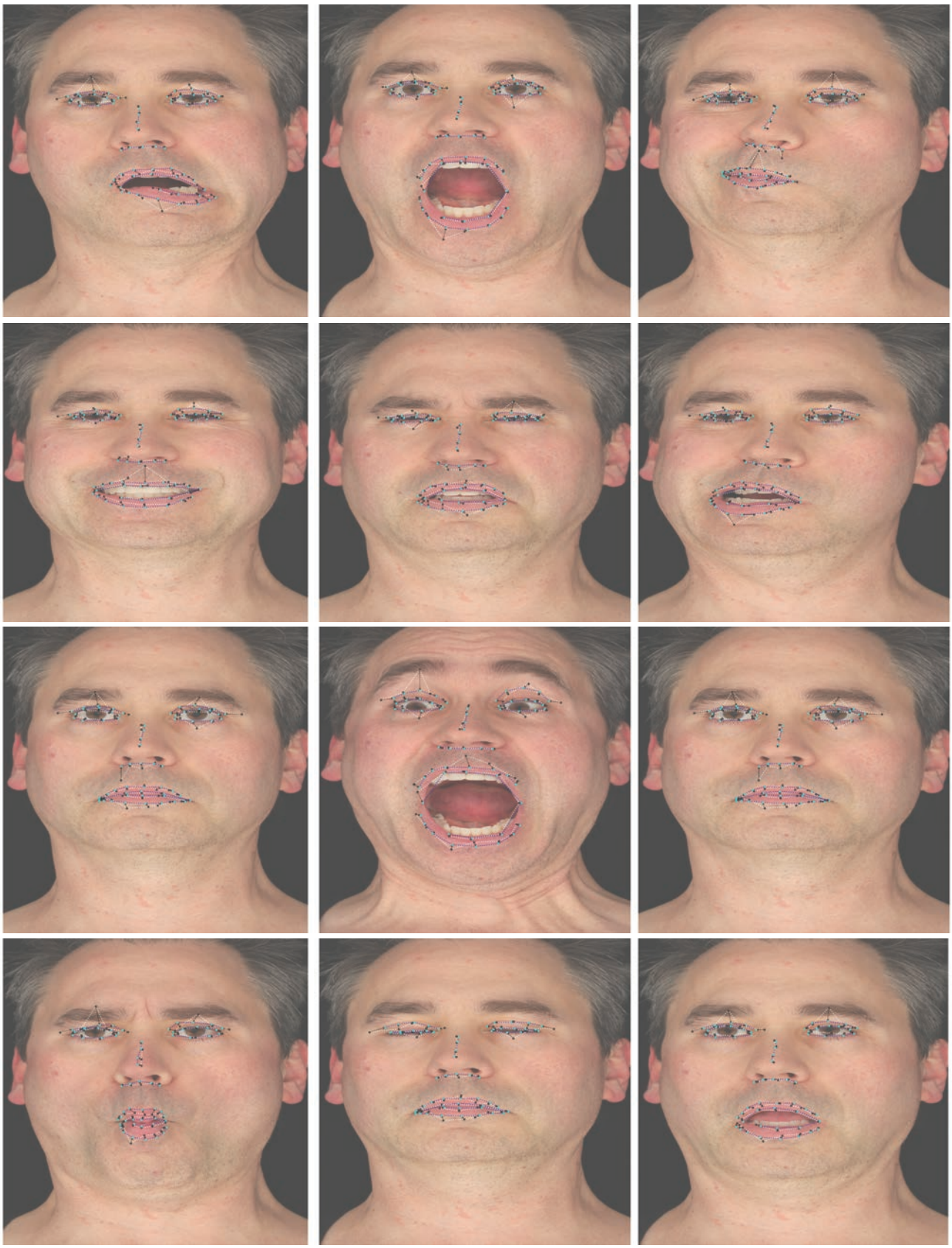

Figure 3. Landmarks shown in black and cyan correspond to predictions by our global and regional models respectively. These results highlight the effectiveness of the regional models on making highly precise predictions.

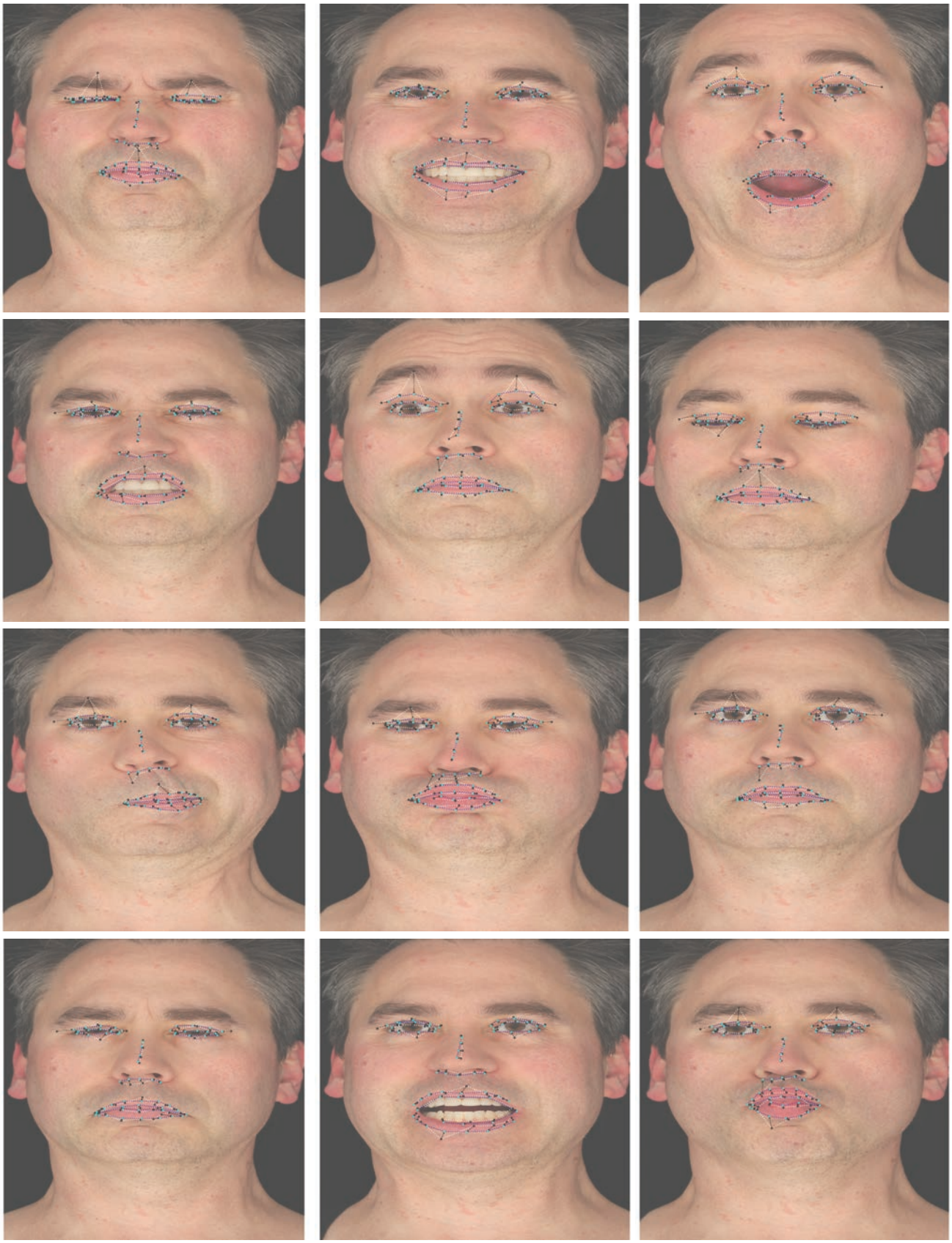

Figure 4. This is a continuation of Fig. 3, showing the corrections made by the regional modules on the remaining expressions of the same subject.

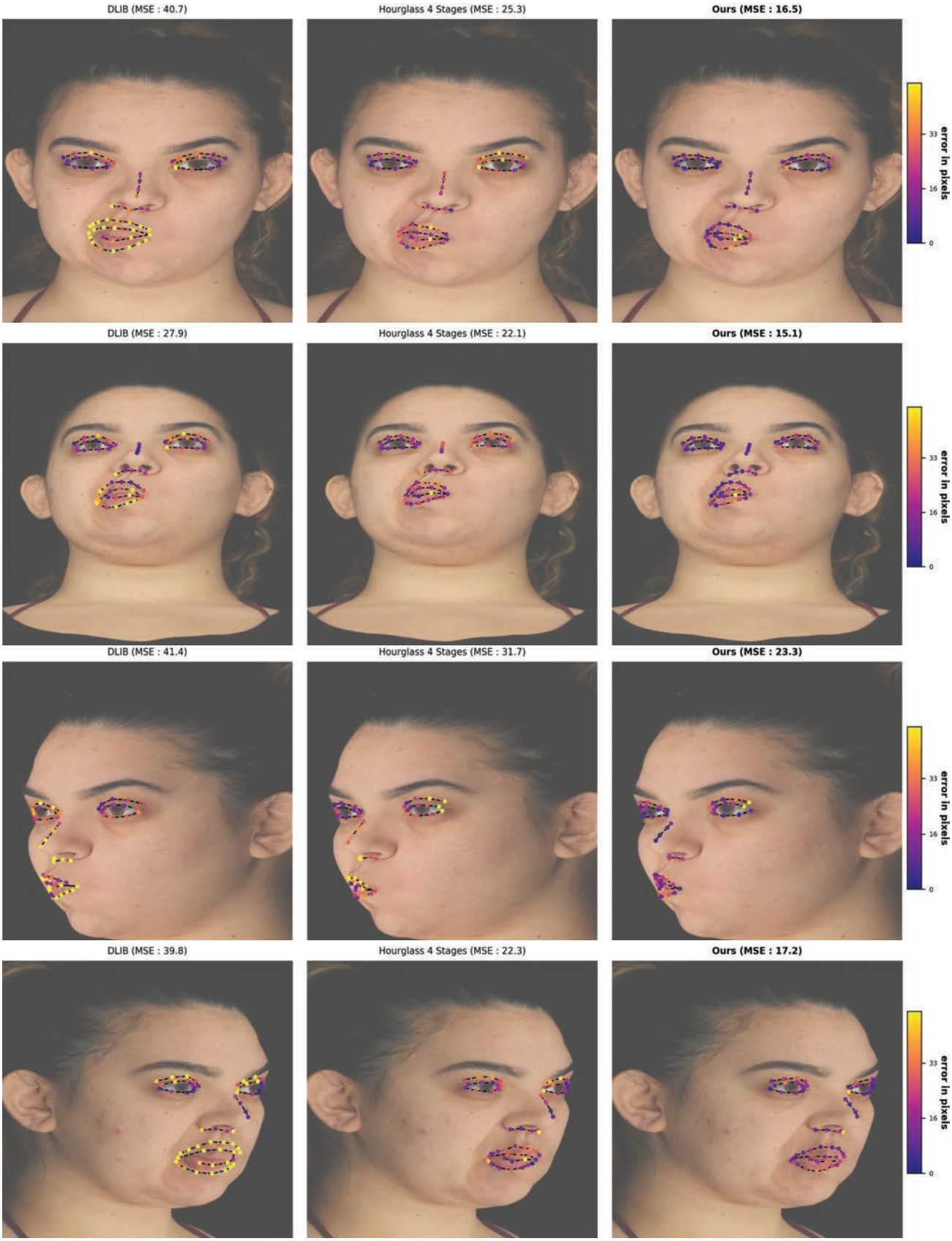

Figure 5. We compare the predictions of DLIB [4], a 4 stage hourglass [1] and our method, for 4 different viewpoints of a test subject. Our method obtains the lowest mean squared error in every viewpoint.

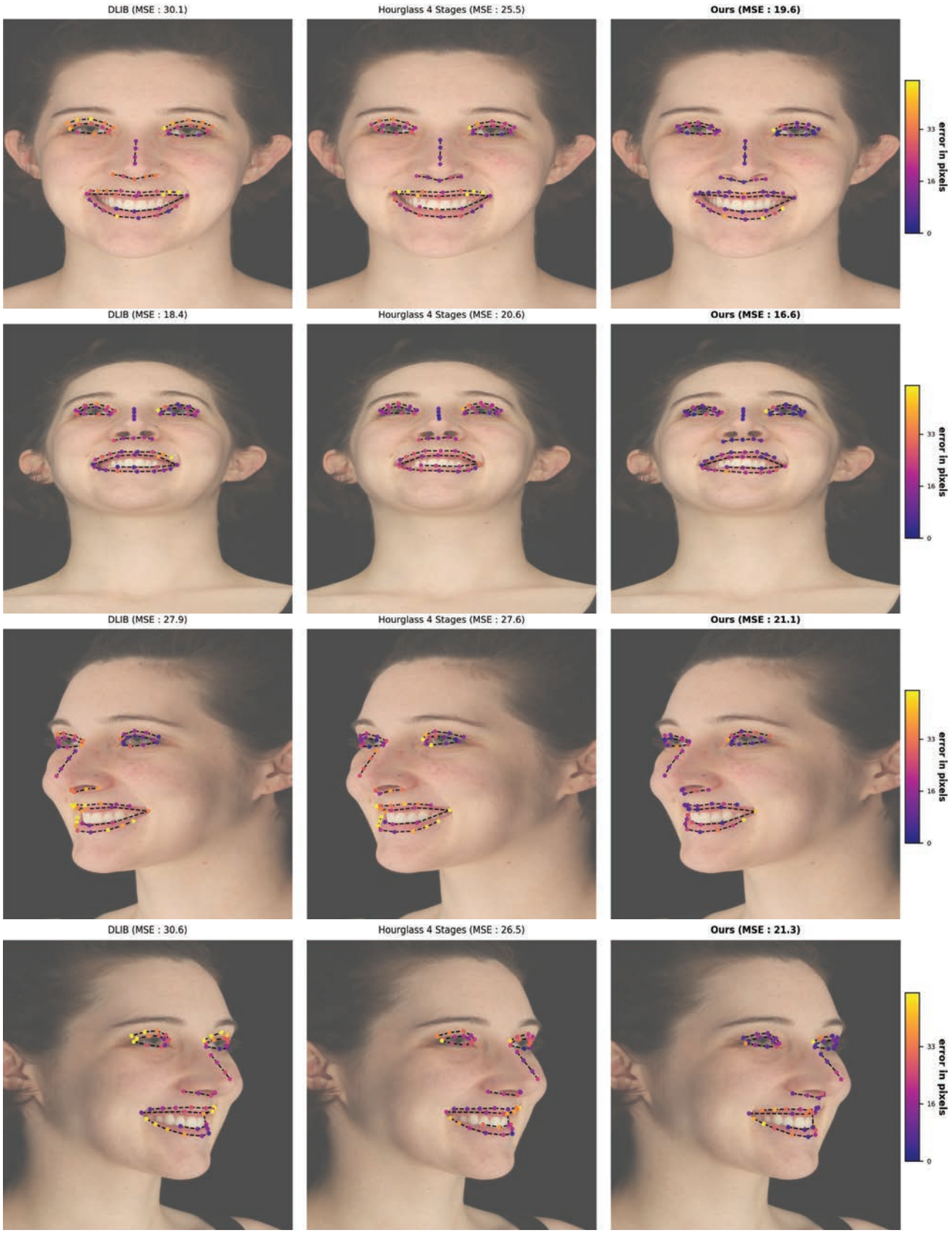

Figure 6. In this figure, we show qualitative comparisons for 4 different viewpoints of a different test subject. Like Fig. 5, our method results in the lowest mean squared error.

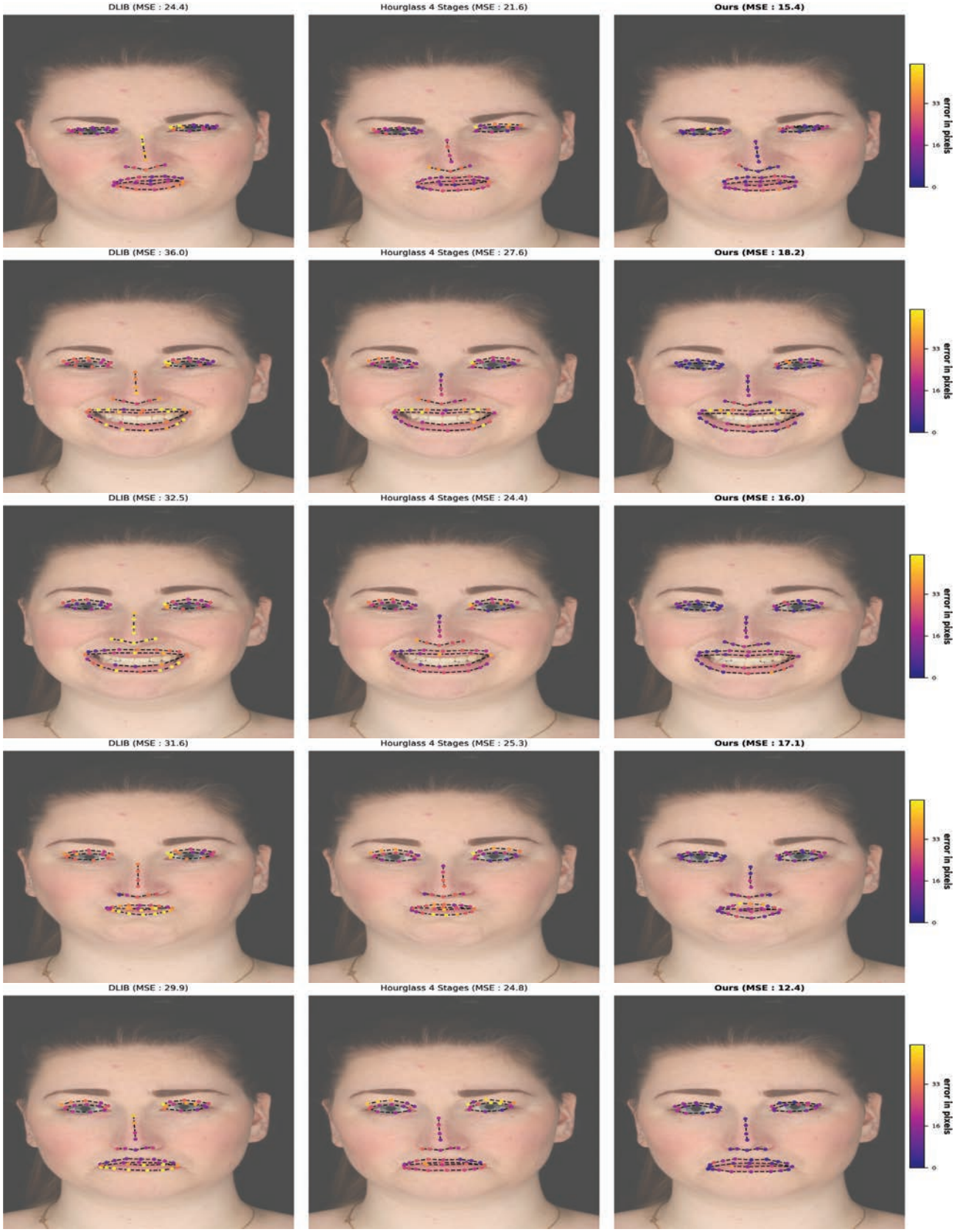

Figure 7. In this figure, we show qualitative comparisons for different expressions of a test subject.

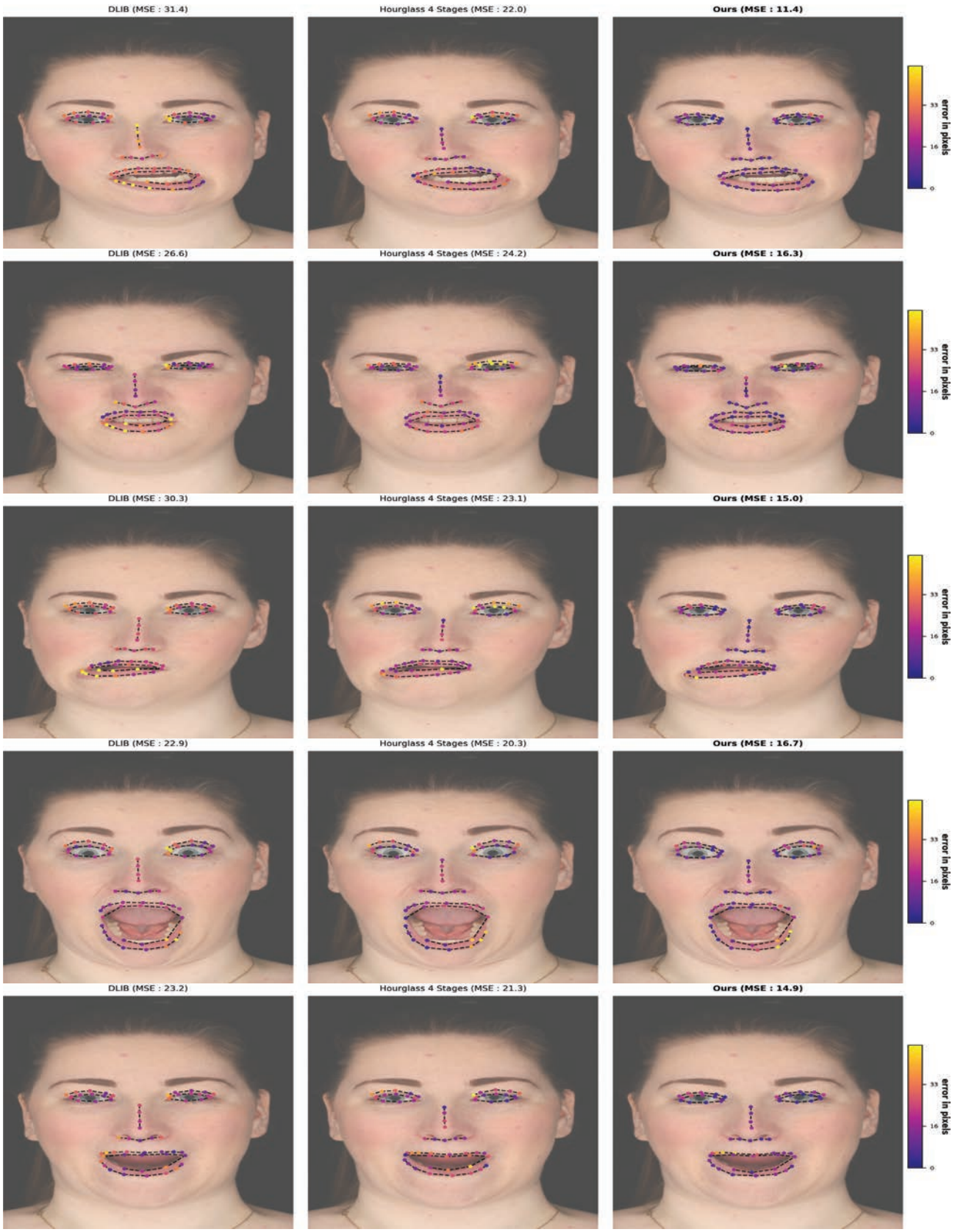

Figure 8. This is continuation of Fig. 7 showing qualitative comparisons across expressions.

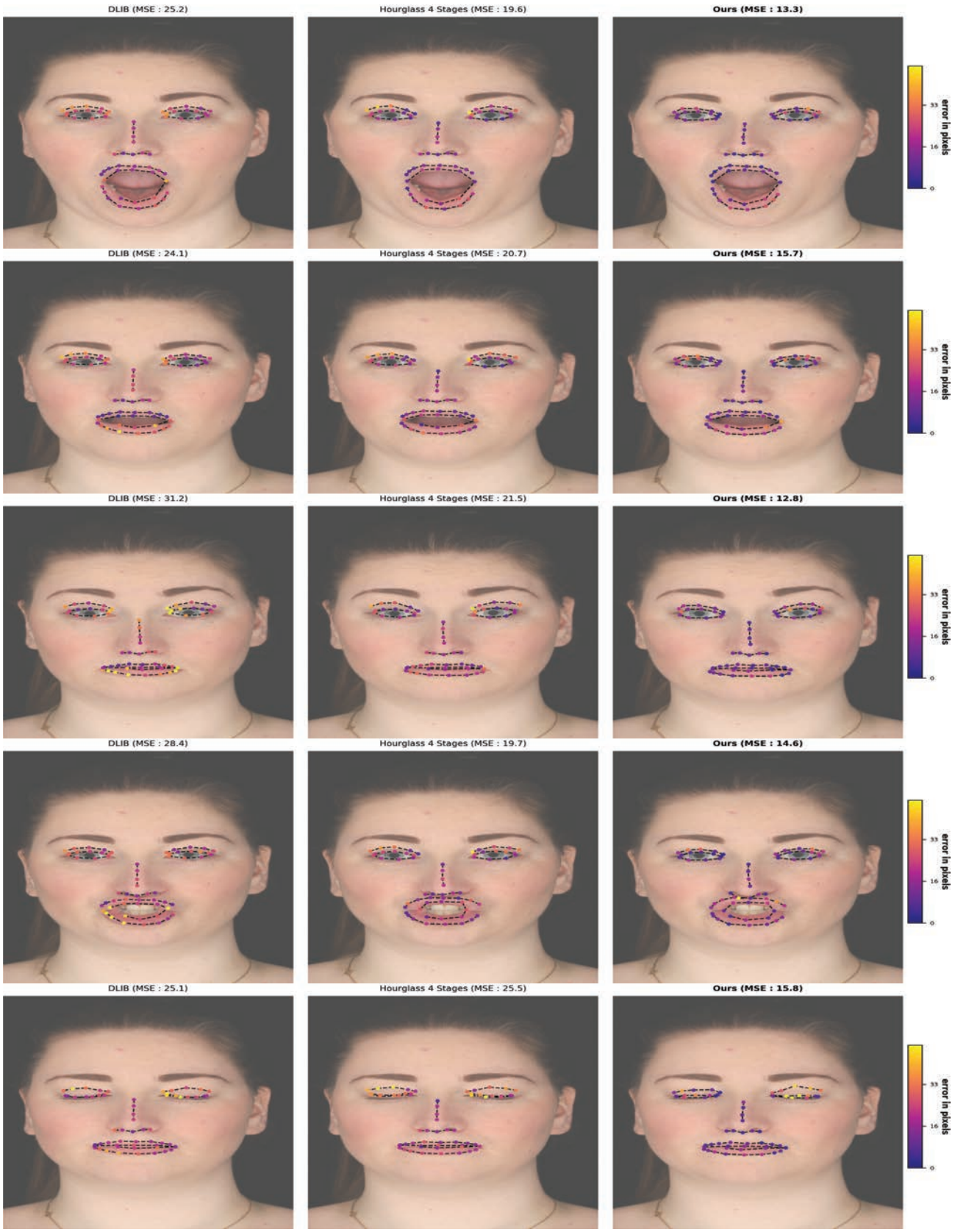

Figure 9. This is continuation of Fig. 7 showing qualitative comparisons across expressions.

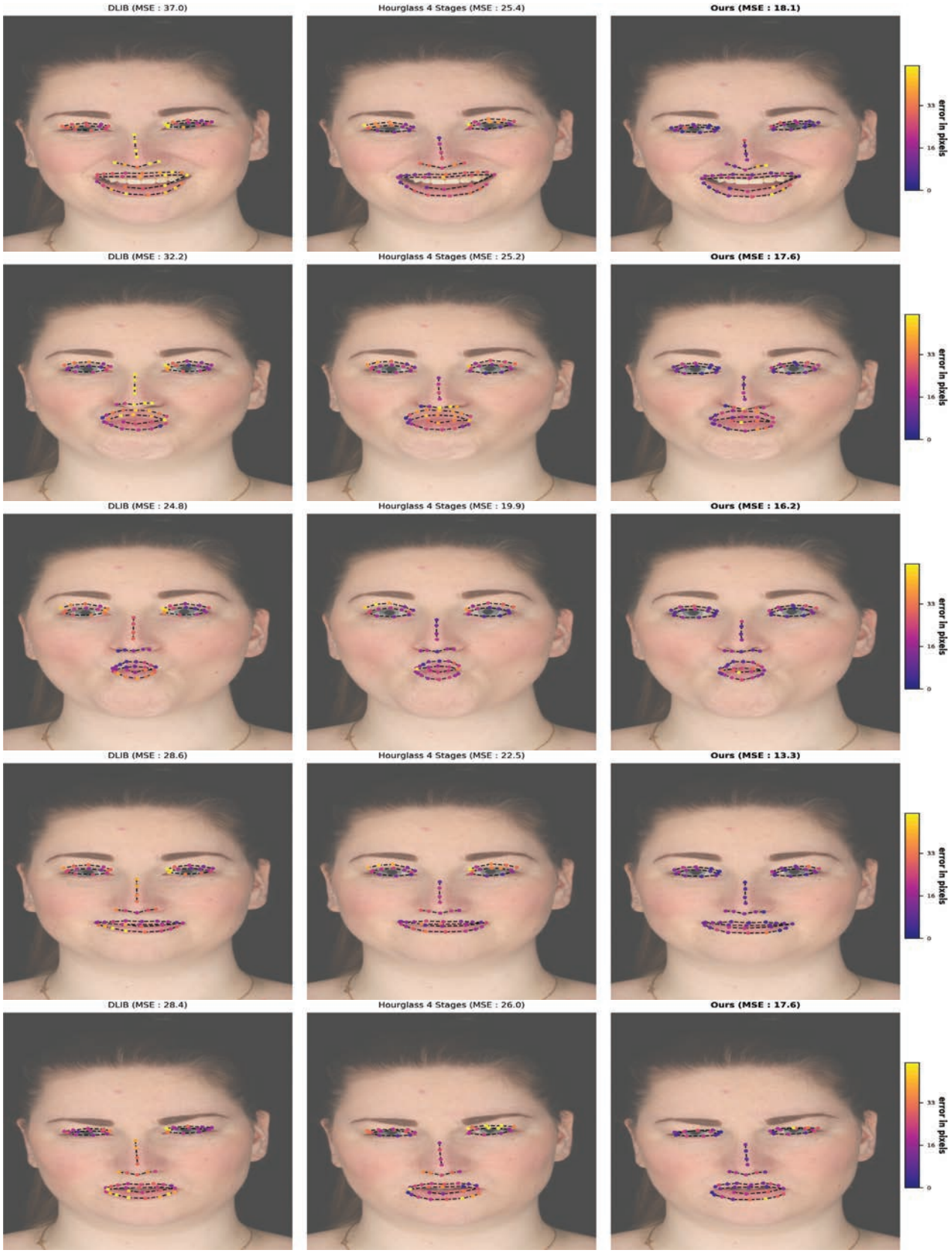

Figure 10. This is continuation of Fig. 7 showing qualitative comparisons across expressions.

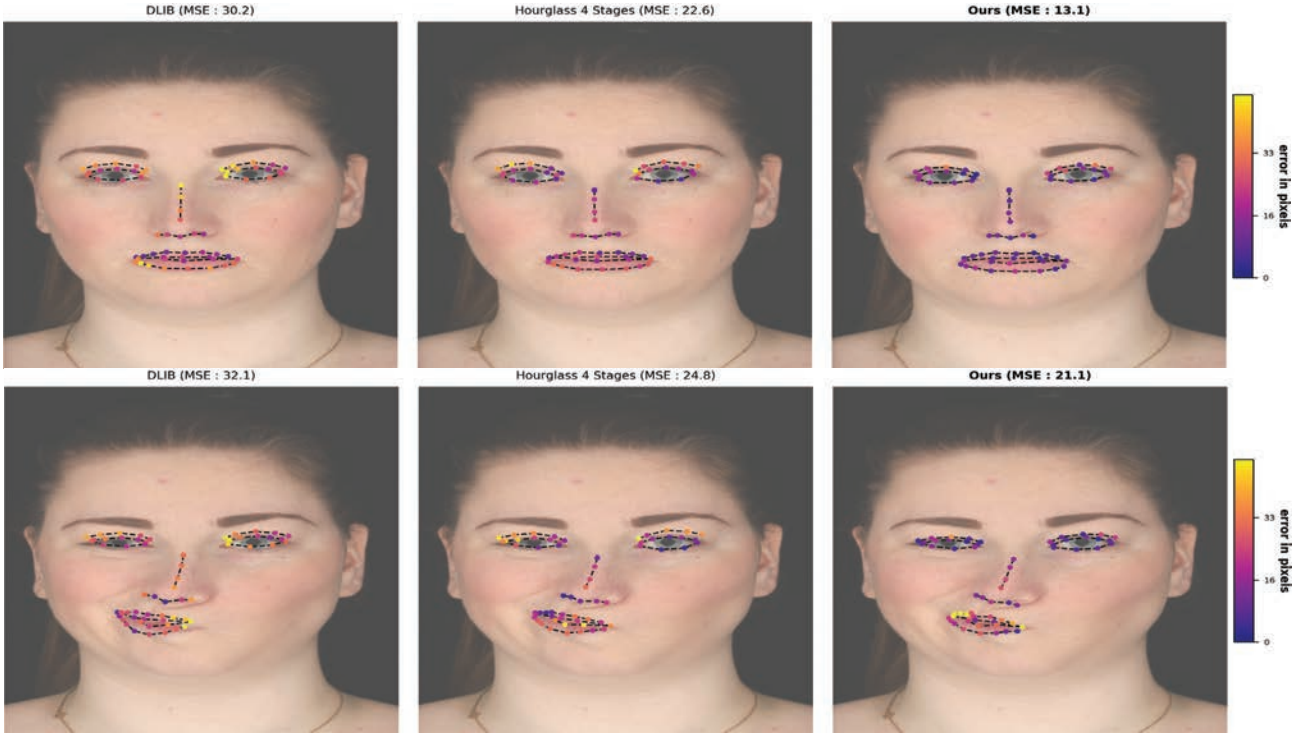

Figure 11. This is continuation of Fig. 7 showing qualitative comparisons across expressions.