Supplementary Material of DenseReg: Fully Convolutional Dense Shape Regression In-the-Wild

1. Mixture of Experts and Quantized Regression

In the paper, we indicate that the proposed quantized regression can interpreted as a 'hard' version of the well known mixture of regression experts [4] model. Herein, we further elaborate upon this interpretation. Similar to the paper, in the mathematical formulation we use the regression of the horizontal field u^h only.

At the testing setup, the final estimate of u^h is computed from the regressed quantized ($\hat{q}^h$) and residual ($\hat{r}^h$) fields as:

$$\hat{u}^h = \hat{q}^h d + \hat{r}^h_{\hat{q}^h}, \tag{1}$$

where $\hat{q}^h$ is modeled using a categorical distribution and is trained using softmax followed by cross entropy loss. This reconstruction can also be seen as:

$$\hat{u}^h = \sum_{i=0}^{K-1} 1_{(\hat{q}^h=i)}(i \cdot d + \hat{r}^h_i), \tag{2}$$

where $(i \cdot d + \hat{r}^h_i)$ is the reconstruction by the i_{th} regressor and $1_{(\hat{q}^h=i)}$ is an indicator function, determining when the i_{th} regressor is active. Note that $i \cdot d$ is the value of $\hat{q}^h$, where i_{th} regressor is active.

Instead of this hard quantization, one can use a soft-quantization using the softmax function as:

$$\hat{u}^h = \sum_{i=0}^{K-1} \left(\frac{e^{f_i^{q^h}}}{\sum_j e^{f_j^{q^h}}} \right) (i \cdot d + \hat{r}^h_i), \tag{3}$$

where f^{q^h} is the output of the CNN branch trained for the quantized ($\hat{q}^h$) field. Notice that this is the *mixture of experts* model[4], where the soft-quantization is analogous to the output of the gating network. It is straightforward to change our model accordingly: shifting each $\hat{r}^h_i$ by adding $(i \cdot d)$ to the bias terms of the corresponding 1×1 convolutional layer and weighting each 'locally trained regressor' output by the softmax function and summing up. Since the parameters of the adapted network are not exactly optimized for this new soft-quantized model, we resort to end-to-end training.

We experiment with fine-tuning the mixture of experts network initialized with the parameters of our hard-quantized network with local regressors. The results are presented in Fig.1. After the fine-tuning, the mixture of experts model performs as well as the quantized regression. Since no significant improvement in regression performance is observed, we have not performed any experiments related to facial analysis with this architecture. We consider that this differentiable representation could be more useful for instance as a spatial transformer network [3], where the deformation field needs to be differentiable.

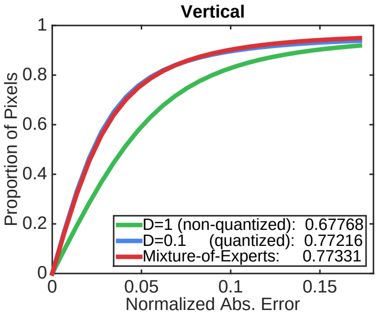


Figure 1: Cumulative Error Distribution of absolute errors normalized by the interocular distance on the deformation-free coordinate system for mixture-of-expers, quantized and non-quantized regression approaches. The mixture of experts model(with no hard quantization) works as well as the quantized model.

2. On the Effect of Label Granularity

In the proposed quantized regression framework, the main motivation is to make use of the robustness and power of CNN's in estimating categorically distributed variables, already demonstrated in the context of semantic segmentation [1]. To quantize our continuous signal, there is a single design parameter, K, which is the number of quantized regions. It determines the quantization step size $d = \frac{1}{K}$. As K increases, the granularity of the discrete label tesselation is increasing.

To analyze the effect of label-space granularity, we perform an experiment by only focusing on the discrete part of the problem. Firstly, we train a set of networks with varying granularities. Then, we evaluate the performance of these networks for classification tasks with different number of classes. Specifically, we measure the segmentation accuracy if (a) the label map is delivered at a level of granularity K and (b) the CNN is trained at a level of granularity M, where $M > K$, and the prediction is obtained by coarsening the CNN's outputs. Our aim in this experiment is to examine what happens as M increases, i.e. as the classifier is trained with increasingly refined labels. The results for the experiment are presented in Table. 1. The evaluation measure used is the average Intersection-over-Union(IoU) for all foreground classes. The best performing network for each granularity is given in boldface.

Table 1: Segmentation performance of classifiers trained with varying number of labels (M). Each classifier is evaluated for classification problems with different number of labels (K). Coarser classification results are obtained by fusing regions.

	M=1	M=2	M=4	M=8	M=16	M=32	M=64
K=1 (fg/bg)	90.43	90.53	90.57	90.67	**90.72**	90.52	89.62
K=2		87.58	87.88	88.20	**88.43**	88.33	87.55
K=4			83.23	83.87	84.32	**84.50**	83.98
K=8				77.09	77.83	**78.30**	78.09
K=16					67.76	68.55	**68.62**
K=32						55.58	**56.08**
K=64							**40.63**

The results indicate that refining the label space simplifies the classification problem by giving the classifier a set of better-defined sub-problems. However, beyond a level of increased refinement we witness a drop in performance, presumably due to smaller amount of training data per class and increased sensitivity to imperfections in the ground-truth. This experiment is highly motivating for our method as we are trying to strike a balance between accuracy, by breaking down the problem into sub-problems through the tesselation procedure, and avoiding overfitting by stopping the refinement after a certain level and turning to the inherently continuous regression formulation for the final refinement.

3. Experiments

In this section, we provide further qualitative and quantitative results that could not be presented in the paper due to space constraints.

3.1. Monocular Depth Estimation

The fitted template shapes also provide the depth from the image plane. We transfer this information to the visible pixels on the image using the same z-buffering operation used for the deformation-free coordinates (detailed in Sec. **2** of the paper). We adopt this as an additional supervision signal: $Z \in [0, 1]$ and add another branch to our network to estimate the depth along with the deformation-free coordinates. To our knowledge, there is no existing results in literature that would allow a quantitative comparison. We are providing example reconstructions using estimated monocular depth fields at Fig.2. We observe that this additional branch does not affect the performance of other branches and adds little to the complexity, since it is just a 1x1 convolution layer after the final shared convolutional layer.

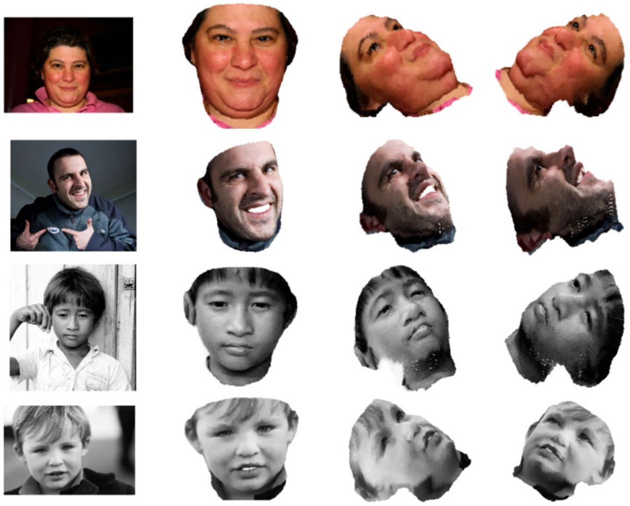


Figure 2: Exemplar 3D renders obtained using estimated depth values.

3.2. Ear Shape Regression

The deformation-free space for the ear shape template is visualized in Fig. 3. The colouring of the qualitative results that are presented in the paper and this supplementary materials document are generated using these coordinates. On Table.2, we provide failure rates and the Area Under Curve(AUC) measures based on the CED curve of the human ear landmark localization experiment, which were not provided in the paper due to space constraints. Further qualitative examples for regressed and ground-truth deformation-free ear coordinates are provided in Fig. 4.

3.3. Quantitative Analysis on Landmark Localization on Static Images and Videos

Static Images: In the paper, we present (Fig. 8, bottom) self-evaluations to compare the *quality of initialization provided to deformable models* by DenseReg and two other standard face detection techniques (HOG-SVM [5], DPM [7]). On Table. 3, we provide quantitative measures to acompany the CED curve provided in the paper: mean, standart deviation, median, mean absolute deviation and

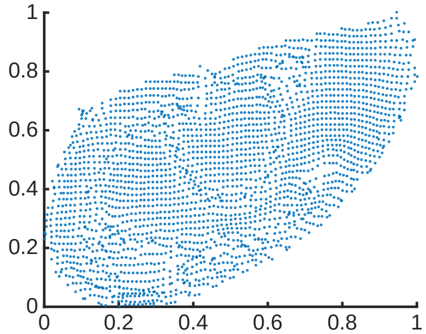


Figure 3: Deformation-free space for the template ear shape.

Method	*AUC*	*Failure Rate (%)*
DenseReg + MDM	**0.4842**	**0.98**
DenseReg	0.4150	1.96
DenseReg + AAM	0.4263	0.98
DPM + MDM	0.4160	15.69
DPM + AAM	0.3283	22.55

Table 2: Landmark localization results on human ear using 55 points. Accuracy is reported as the Area Under the Curve (AUC) and the Failure Rate of the Cumulative Error Distribution of the normalized RMS point-to-point error.

maximum normalized point-to-point error, followed by area under curve(AUC) and failure rate.

Deformable Face Tracking: The 300VW challenge [8, 2] benchmark consists of 114 videos that are separated into three categories: *(i)* Videos captured in well-lit environments without severe occlusions, *(ii)* videos captured in unconstrained illumination conditions, and *(iii)* videos captured in totally arbitrary conditions (severe occlusions and extreme illuminations). In the paper, we have provided the results for all of the categories combined. On Figure 5, we provide the CED curves and corresponding AUC and FR values for each category separately. In terms of AUC, the proposed method, *DenseReg+MDM*, is slightly outperformed by Xiao et al. [9] in Category 2, whereas it achieves slightly better performance compared to all participants in Categories 1 and 3. We again highlight that the tracking results are obtained without using the training set of the 300VW dataset and without using any temporal modeling.

3.4. Qualitative Results

Additional qualitative results from Helen testset [6] are provided for the task of semantic segmentation (Fig. 6) and landmark localization (Fig. 7,8).

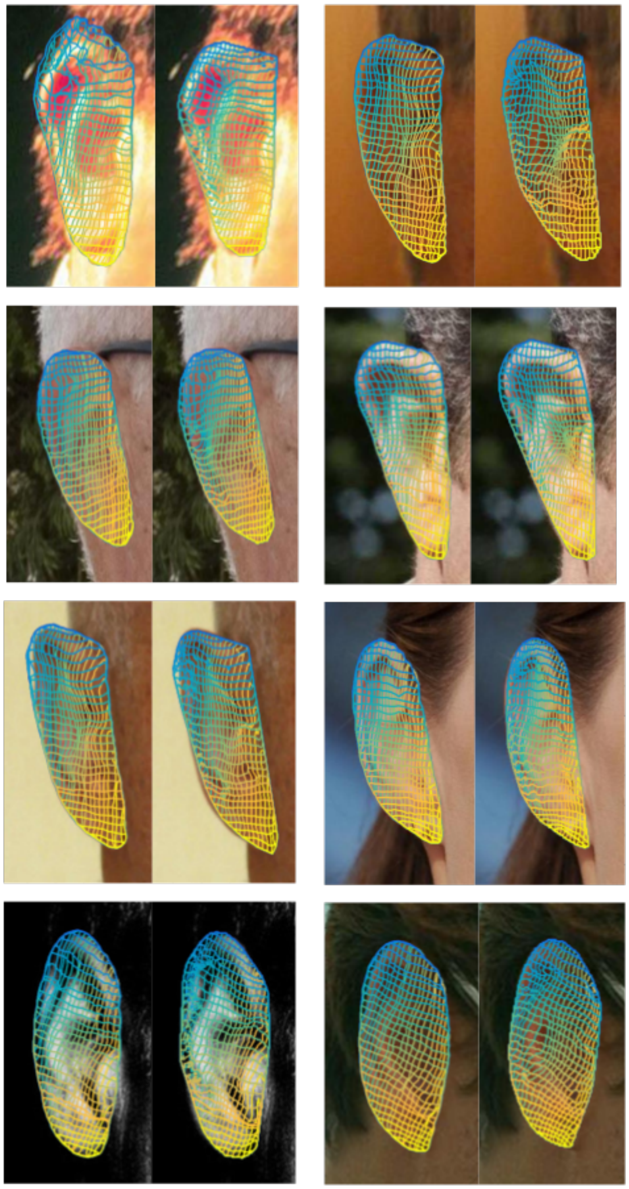


Figure 4: Exemplar pairs of deformation-free coordinates of dense landmarks on human ear. *Left:* Estimated by DenseReg. *Right:* Ground-truth produced by TPS.

References

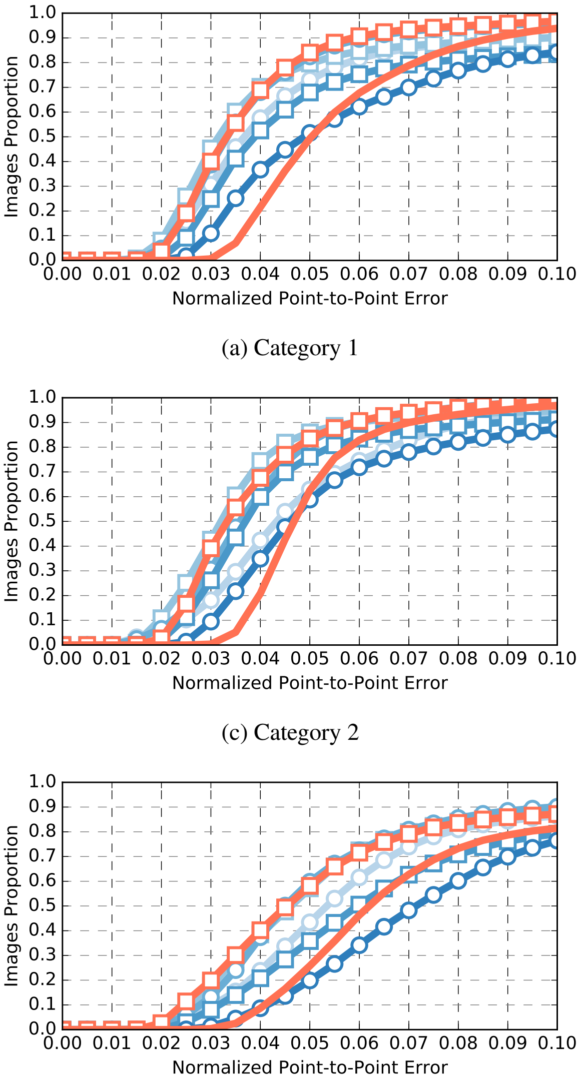
[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 1

[2] G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape. Offline deformable face tracking in arbitrary videos. In *Proceedings of IEEE International Conference on Computer Vision, 300 Videos in the Wild (300-VW): Facial Landmark Tracking in-the-Wild Challenge & Workshop (ICCVW'15)*, Santiago, Chile, December 2015. 3

[3] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Process-*

	Mean	Std	Median	Mad	Max	AUC$_{0.1}$	Failure Rate
DenseReg + MDM	0.0587	0.1636	0.0433	0.0091	3.8061	0.5219	0.0367
DenseReg + AAM	0.0723	0.1926	0.0547	0.0131	4.6353	0.4117	0.0917
DPM + MDM	0.0878	0.1716	0.0567	0.0154	2.8255	0.3768	0.1533
DPM + AAM	0.1371	0.2596	0.0568	0.0182	4.1285	0.3628	0.2317
DenseReg	0.0795	0.2009	0.0606	0.0124	4.8254	0.3605	0.1083
HOG-SVM + MDM	0.1827	0.3997	0.0597	0.0216	3.4132	0.3461	0.2550
HOG-SVM + AAM	0.1396	0.2596	0.0591	0.0185	4.1296	0.3439	0.2400

Table 3: Quantitative results on the test set of the 300W competition (68-points).



(a) Category 1

(c) Category 2

(e) Category 3

	AUC$_{0.1}$	**FR**
DenseReg + MDM	0.6190	0.0328
Deng et al.	0.6136	0.0406
Xiao et al.	0.5973	0.0983
Rajamanoharan et al.	0.5561	0.0988
Wu et al.	0.5065	0.1650
DeepReg	0.4421	0.0612
Unicar et al.	0.4311	0.1580

(b) Category 1

	AUC$_{0.1}$	**FR**
Xiao et al.	0.6300	0.0534
DenseReg + MDM	0.6201	0.0129
Deng et al.	0.6084	0.0144
Wu et al.	0.5590	0.0872
Rajamanoharan et al.	0.5157	0.0601
DeepReg	0.4934	0.0312
Unicar et al.	0.4634	0.1264

(d) Category 2

	AUC$_{0.1}$	**FR**
DenseReg + MDM	0.4857	0.1282
Deng et al.	0.4844	0.0988
Xiao et al.	0.4813	0.1192
Rajamanoharan et al.	0.4227	0.1359
Wu et al.	0.3646	0.2078
DeepReg	0.3362	0.1853
Unicar et al.	0.2801	0.2357

(f) Category 3

Figure 5: Deformable tracking results against the state-of-the-art on the 300VW testing dataset using 68 points. Accuracy is reported as the Area Under the Curve (AUC) and the Failure Rate of the Cumulative Error Distribution of the RMS error normalized with interocular distance.

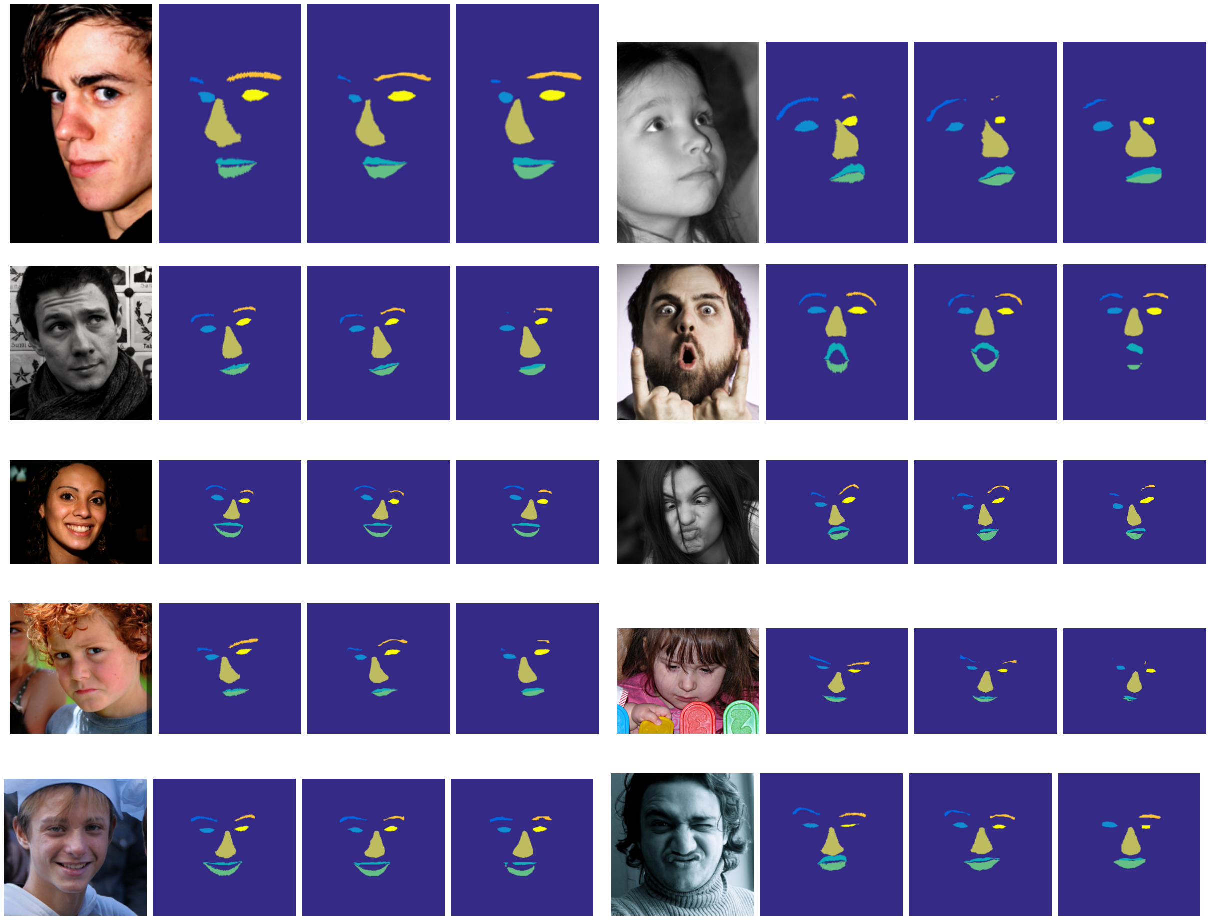


Figure 6: Exemplar semantic segmentation results. *Left:* Ground-truth. *Center:* DenseReg. *Right:* DeepLab-v2.

ing Systems, pages 2017–2025, 2015. 1

[4] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994. 1

[5] D. E. King. Max-margin object detection. *arXiv preprint arXiv:1502.00046*, 2015. 2

[6] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, pages 679–692. Springer, 2012. 3

[7] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *European Conference on Computer Vision*, pages 720–735. Springer, 2014. 2

[8] J. Shen, S. Zafeiriou, G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of IEEE International Conference on Computer Vision, 300 Videos in the Wild (300-VW): Facial Landmark Tracking in-the-Wild Challenge & Workshop (ICCVW'15)*, December 2015. 3

[9] S. Xiao, S. Yan, and A. A. Kassim. Facial landmark detection via progressive initialization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 33–40, 2015. 3

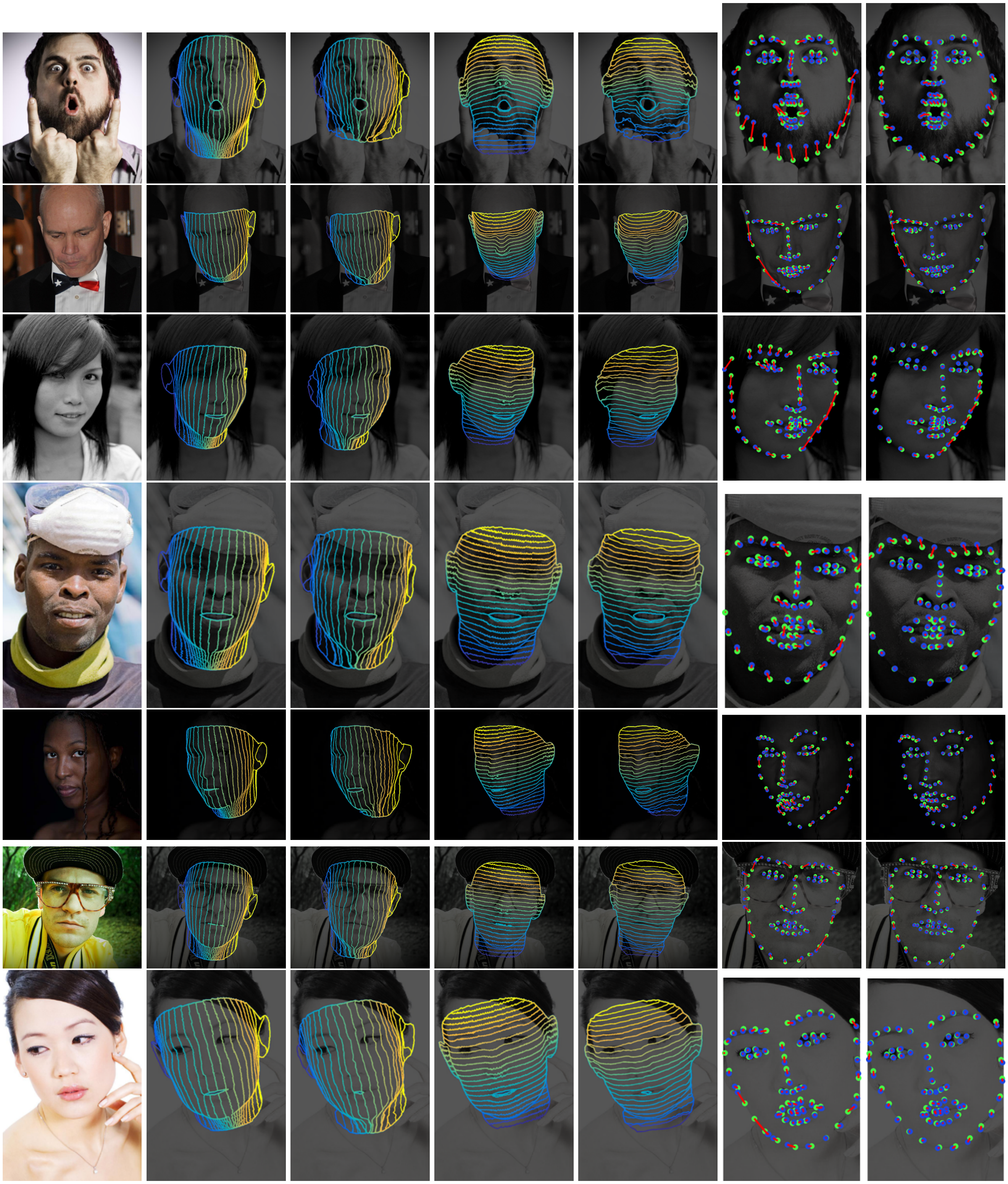


Figure 7: Qualitative Results. From left to right: Original image, ground-truth horizontal coordinates(u^h), estimated horizontal coordinates($\hat{u}^h$) , ground-truth vertical coordinates(u^v), estimated vertical coordinates($\hat{u}^v$) , Landmarks for DenseReg, Landmarks for DenseReg+MDM. Estimated landmarks(blue), ground-truth(green), lines between estimated and ground-truth landmarks(red).

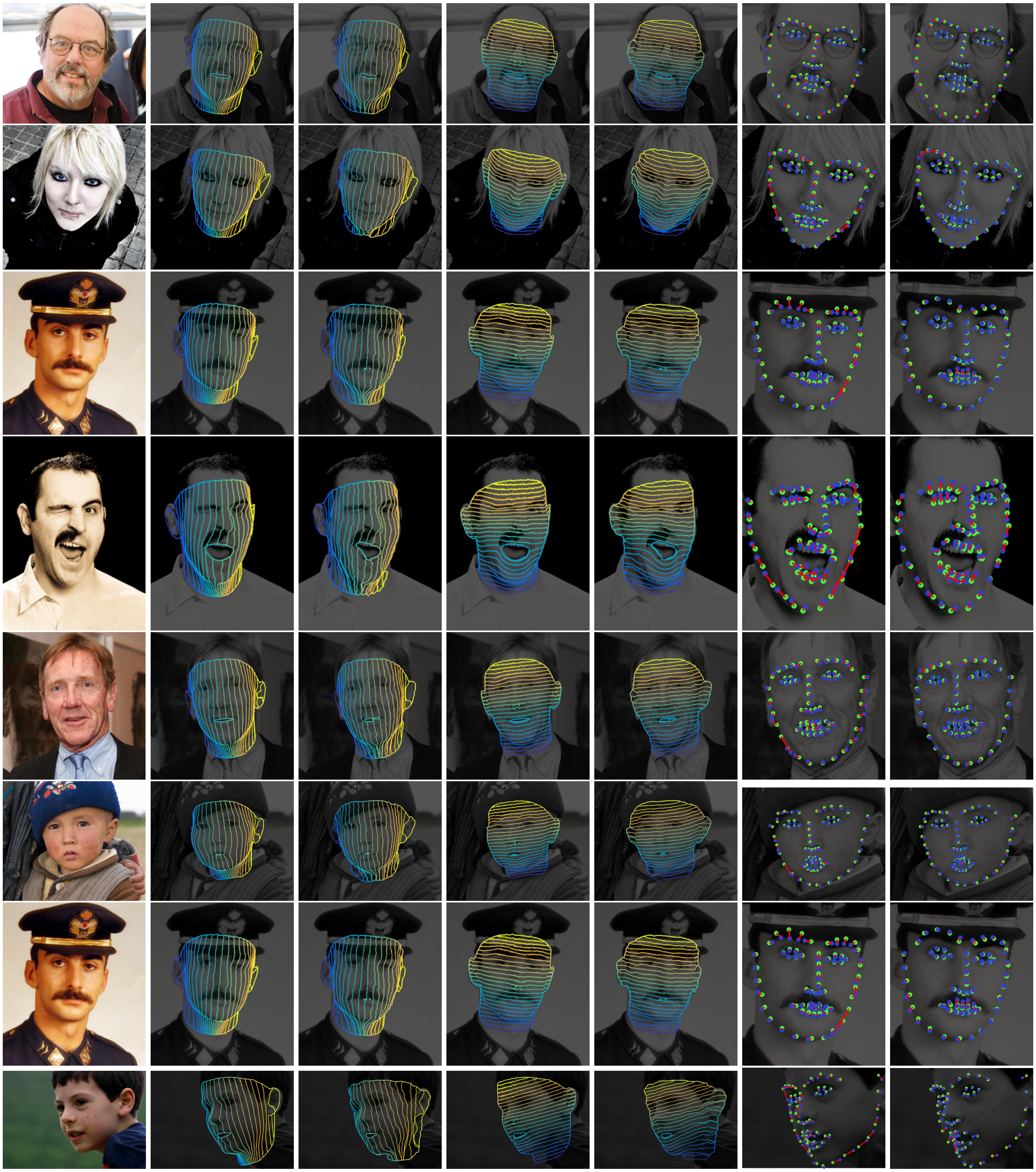


Figure 8: Qualitative Results. From left to right: Original image, ground-truth horizontal coordinates(u^h), estimated horizontal coordinates($\hat{u}^h$) , ground-truth vertical coordinates(u^v), estimated vertical coordinates($\hat{u}^v$) , Landmarks for DenseReg, Landmarks for DenseReg+MDM. Estimated landmarks(blue), ground-truth(green), lines between estimated and ground-truth landmarks(red).