Supplementary information

This is a supplementary material caused restrict by the length of the main text.

1 More Ablation Study Results

We conducted comparative tests on the M4Net$_{base}$.

Fig. S1 shows the ablation study results of different multi-head weights. Dynamic means the weights from a FC layer, according the experiments, finally we choose a set of static parameter.

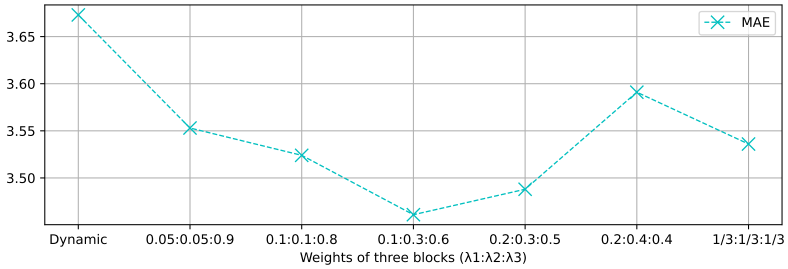

Figure S1. The impact of weight on accuracy.

Fig. S2 shows the ablation study results of different activation functions. It can be seen from the line chart and violin chart that the ReLU series(ReLU, ELU, and RReLu) activation function have slightly better convergence speed, train loss and volatility than Sigmoid, but there is no significant difference in the quantitative and qualitative test.

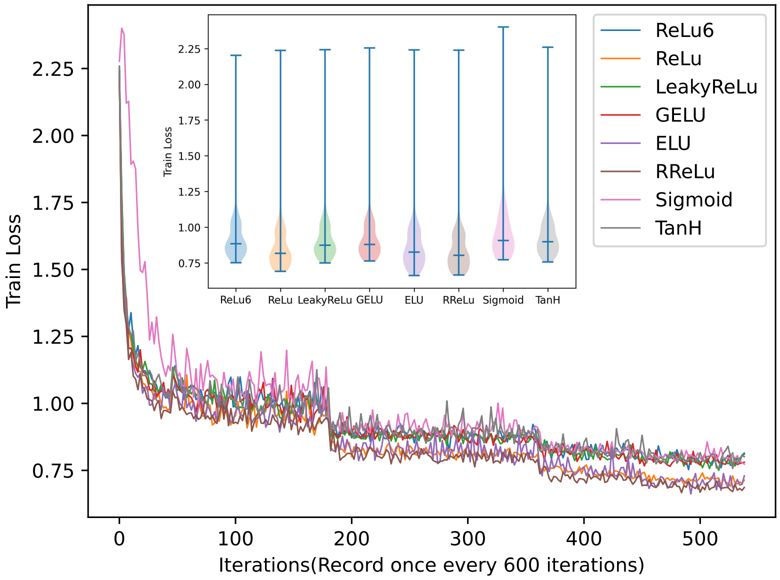

Figure S2. Comparing the training losses of 8 commonly used activation functions.

2 More Qualitative Experiment Results

We provide additional visual results in this section to further demonstrate the effectiveness of this method, testing images from the FOA[1] and UFDD[2].

FOA avatars are collected from various artworks, with pixel features significantly differing from real human faces. As shown in Fig. S3, our method exhibits good robustness, successfully completing tasks such as landmark detection, head pose estimation, and dense face alignment (3D mesh reconstruction). However, this example also reveals limitations of 3DMM-based approach: while the mean face model is derived from real individuals and can adequately fit head poses and facial landmarks, it cannot precisely reconstruct exaggerated cartoon facial features. Consequently, the generated faces lack expressiveness in terms of personalization.

UFDD datasets divided into several sub-datasets by different scenes, following the original partitioning of the datasets, we present partial visual results for seven scenes: focus, haze, illumination and lens, motion, snow and rain, in Figures S4, S5, S6, S7 and S8, respectively. From these results, it is evident that our method performs well in complex lighting conditions and scenes obstructed by rain, snow, or fog. When dealing with scenes containing multiple individuals in a single image, the industry typically adopts two approaches: top-down and bottom-up.

By the way, we follow the mainstream method[3–5] to employ a top-down approach, incorporating a third-party facial detection module for face detection, Specifically, this paper followed SynergyNet[3] use Faceboxs[6]. The cropped facial region information is then fed into our model, enabling multi-person detection without the need for additional training.

References

1. Yaniv, J., Newman, Y. & Shamir, A. The face of art: Landmark detection and geometric style in portraits. In *ACM Transactions on Graphics (Proceedings SIGGRAPH)*, vol. 38, 60:1–60:15 (2019).

2. Nada, H., Sindagi, V., Zhang, H. & Patel, V. M. Pushing the limits of unconstrained face detection: a challenge dataset and baseline results. *arXiv preprint arXiv:1804.10275* (2018).

3. Wu, C. Y., Xu, Q. G., Neumann, U. & Soc, I. C. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *9th International Conference on 3D Vision (3DV)*, International Conference on 3D Vision, 453–463 (Ieee Computer Soc, LOS ALAMITOS, 2021).

4. Li, H., Wang, B., Cheng, Y., Kankanhalli, M. & Tan, R. T. Dsfnet: Dual space fusion network for occlusion-robust 3d dense face alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4531–4540 (2023).

5. Guo, J. *et al.* Towards fast, accurate and stable 3d dense face alignment. In *European Conference on Computer Vision*, 152–168 (Springer, 2020).

6. Zhang, S., Wang, X., Lei, Z. & Li, S. Z. Faceboxes: A cpu real-time and accurate unconstrained face detector. *Neurocomputing* **364**, 297–309 (2019).

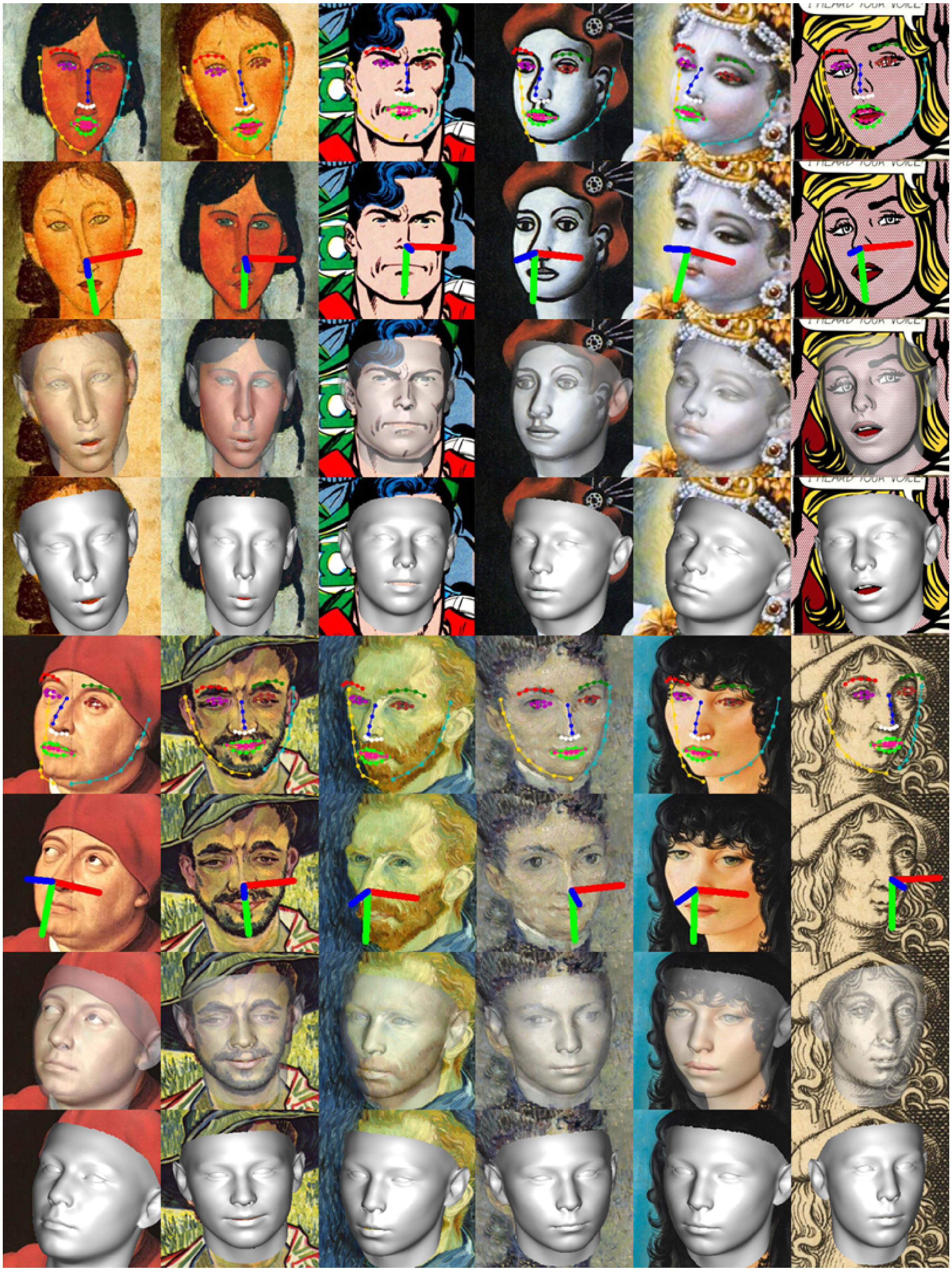

Figure S3. Visualization of 3D face geometry prediction on FOA[1] from M4Net$_{base}$. Row 1-4: Landmarks, Head Pose Estimation, 3D face mesh(translucent), 3D face mesh(solid).

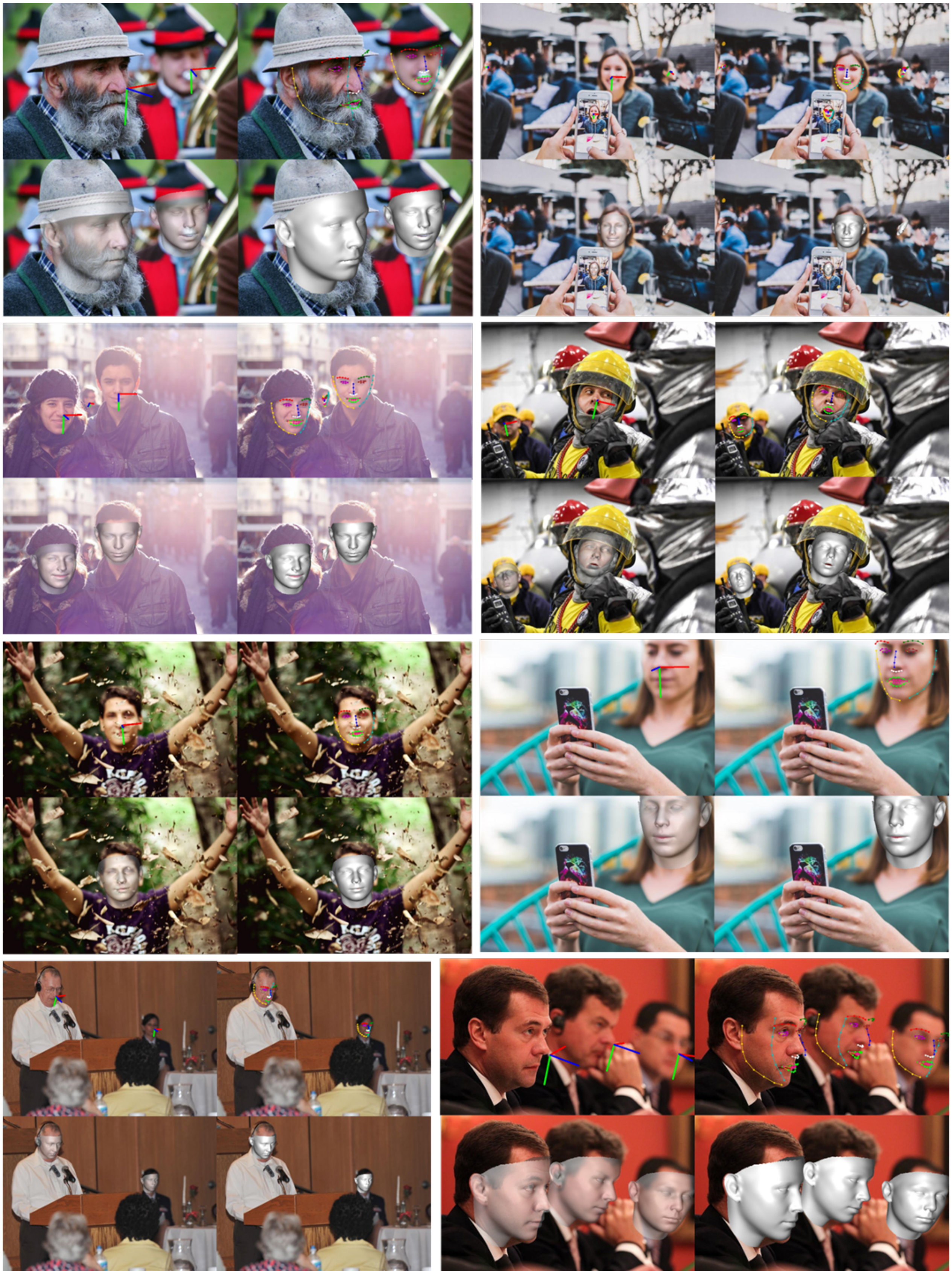

Figure S4. Visualization of 3D face geometry prediction on UFDD(Scene: Focus)[2] from M4Net$_{base}$.

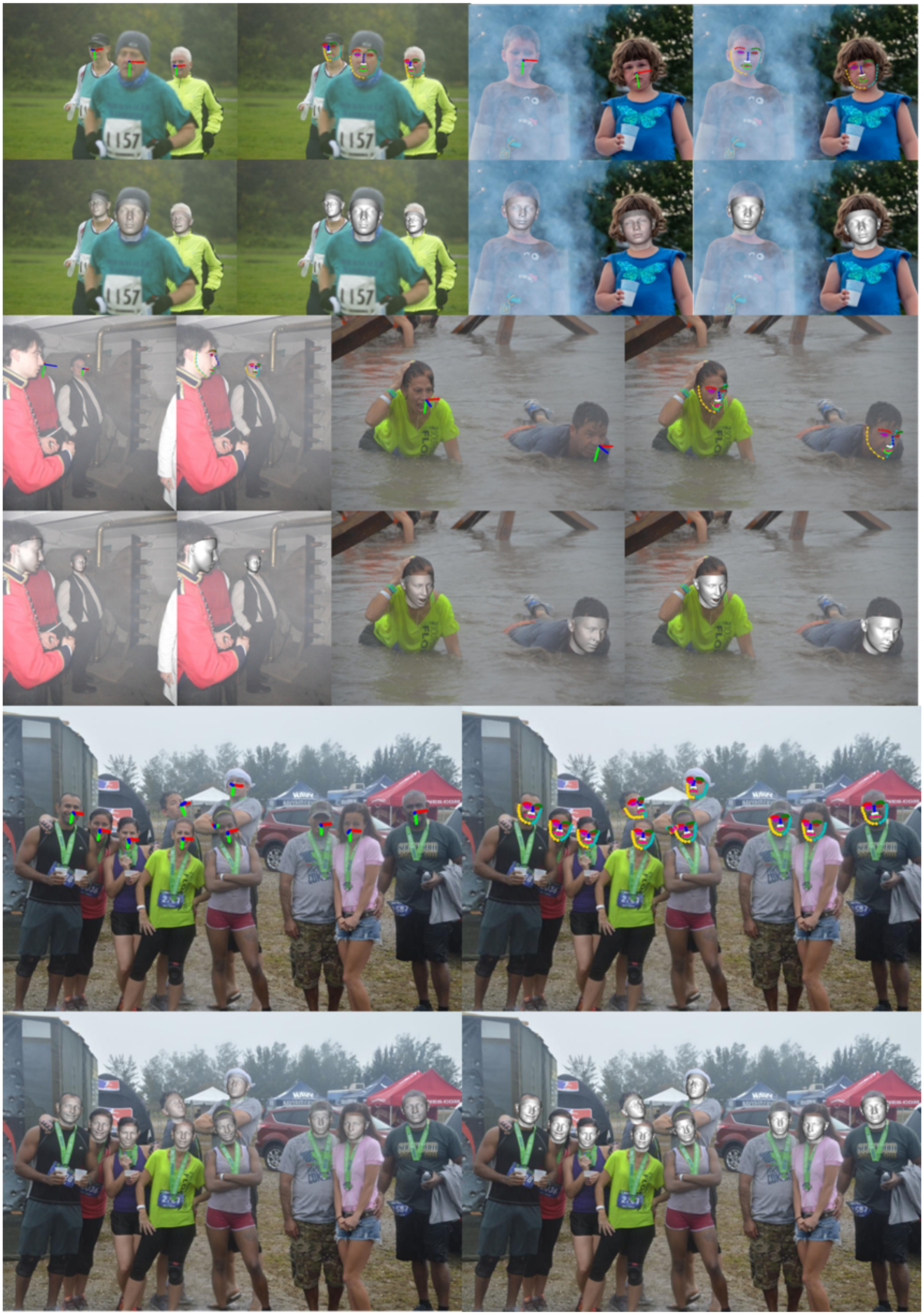

Figure S5. Visualization of 3D face geometry prediction on UFDD(Scene: Haze)[2] from M4Net$_{base}$.

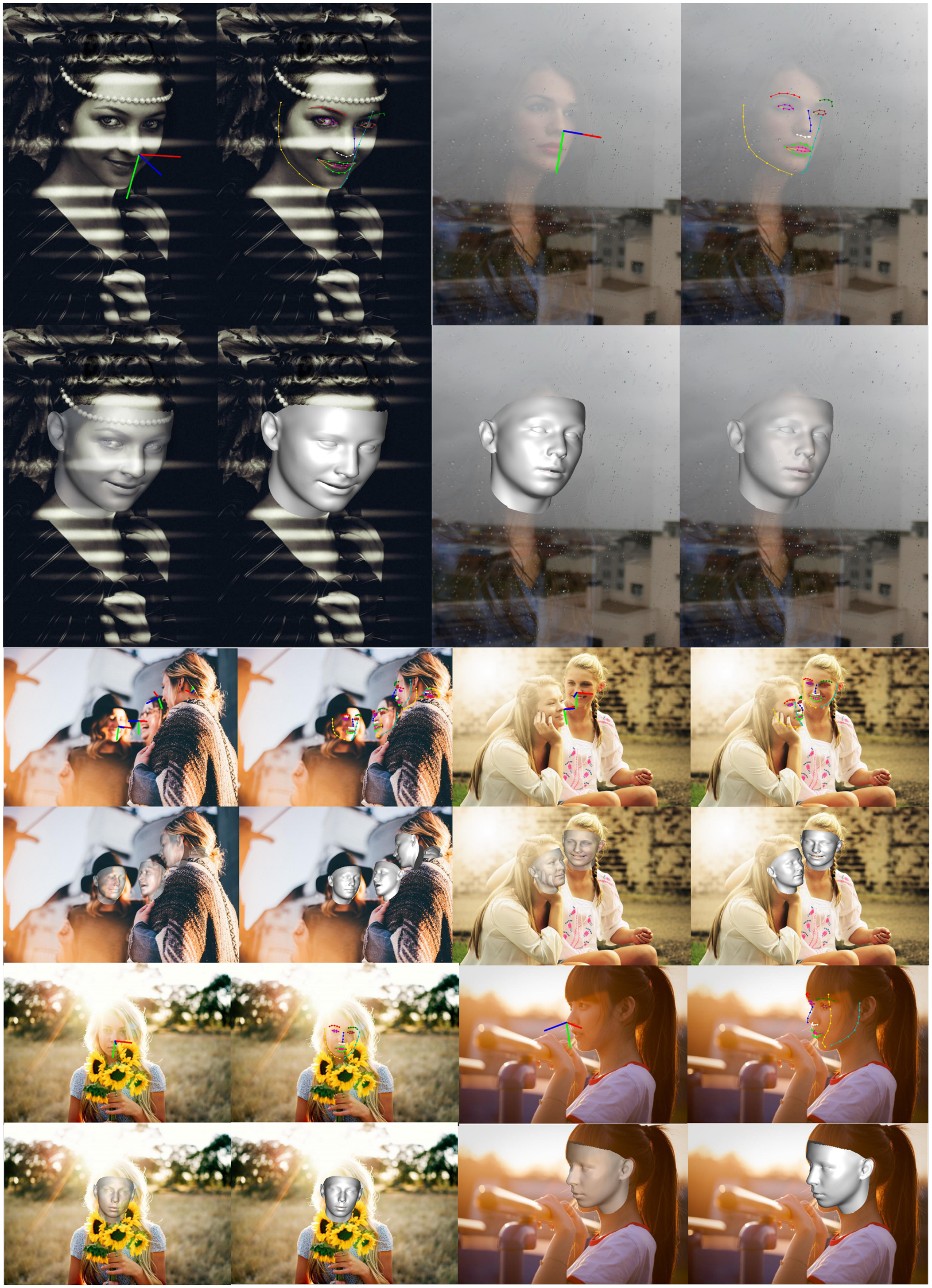

Figure S6. Visualization of 3D face geometry prediction on UFDD(Scene: Illumination and lens)[2] from M4Net$_{base}$.

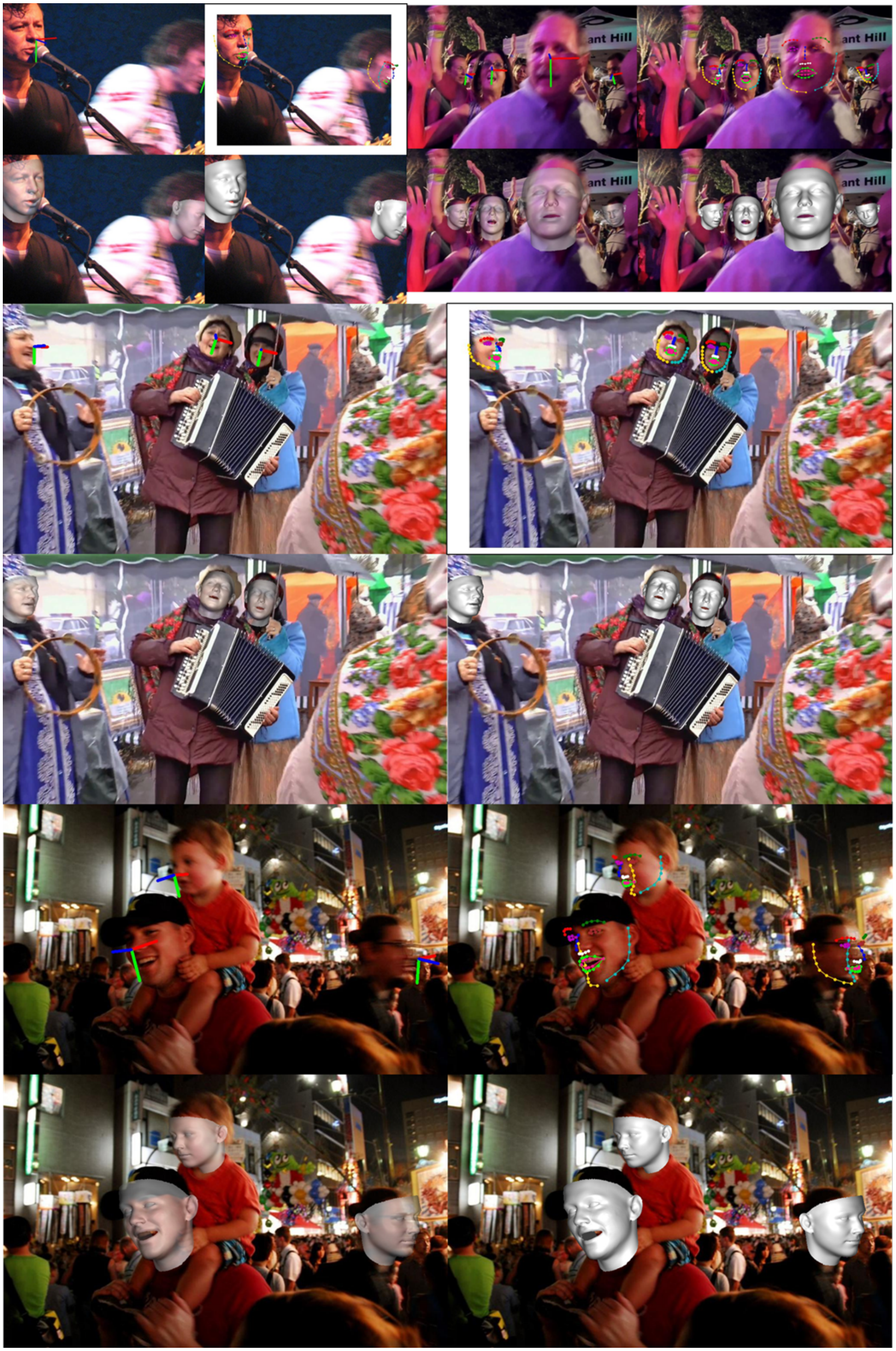

Figure S7. Visualization of 3D face geometry prediction on UFDD(Scene: Motion)[2] from M4Net$_{base}$.

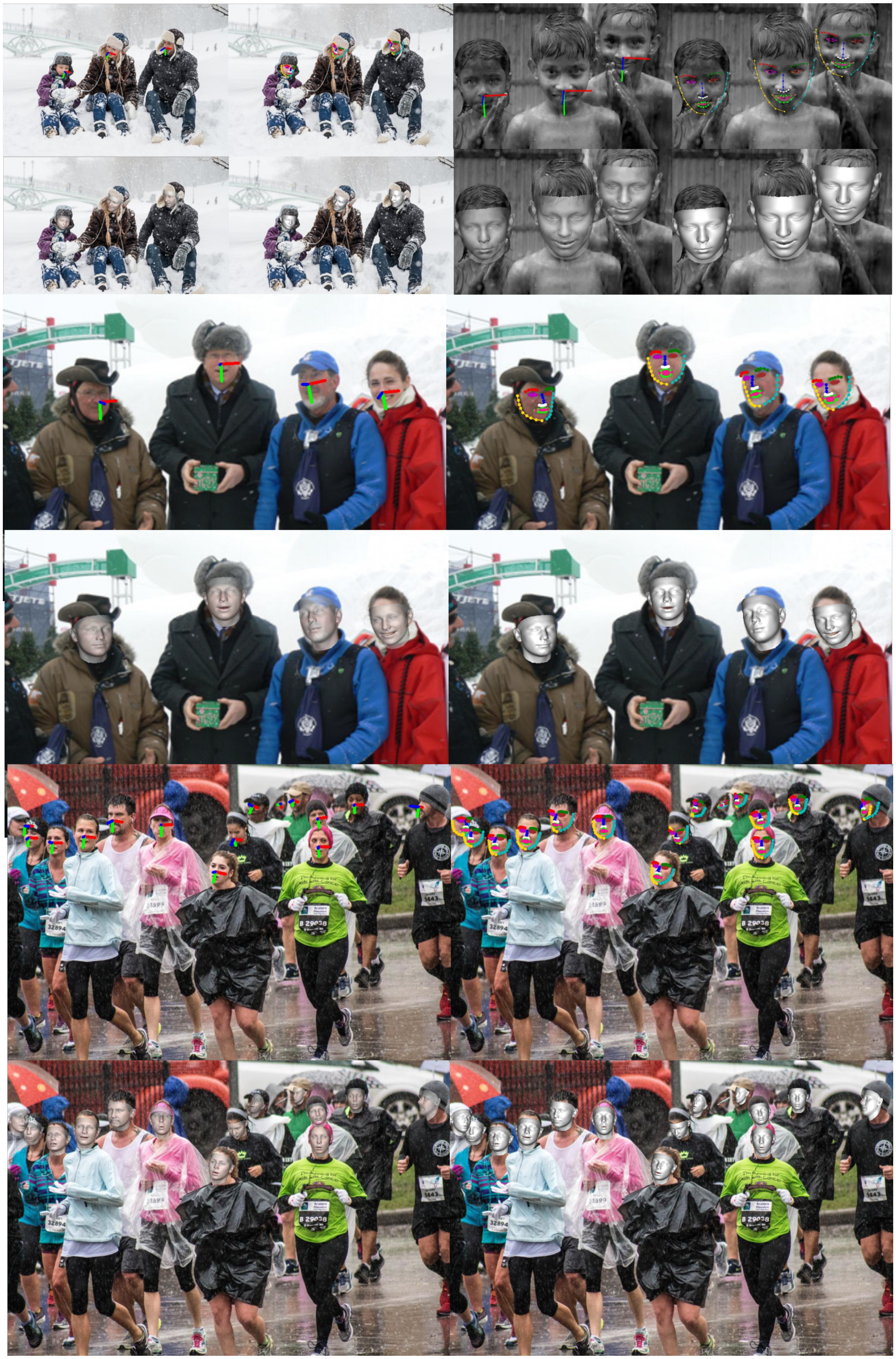

Figure S8. Visualization of 3D face geometry prediction on UFDD(Scene: Snow and Rain)[2] from M4Net$_{base}$.