3FabRec: Fast Few-shot Face alignment by Reconstruction

Supplementary Materials

Björn Browatzki and Christian Wallraven*
Dept. of Artificial Intelligence, Korea University, Seoul
browatbn@korea.ac.kr, wallraven@korea.ac.kr

Figure 1: Randomly-generated faces from the 3FabRec framework (top four rows) together with their predicted landmark confidence heatmaps (bottom four rows).

1. Summary of supplementary materials

In the following experiments, we present further visualizations for randomly-generated faces (Sec. 2), several additional ablation studies on losses, encoding length, and different training setups (Sec. 3), and visualizations for results from few-shot learning on different datasets (Sec. 4).

2. Random faces

Figure 1 shows generated faces from a random sampling of the latent space (top four rows) together with the predicted landmark heatmaps (bottom four rows) using the final architecture from the main paper (trained on VGGFace2 and AffectNet with 256x256px). We note that the faces

	Model	$\mathcal{L}_{rec}$	$\mathcal{L}_{adv}$	$\mathcal{L}_{cs}$	FT	**Global Reconstr.** RMSE	SSIM	**Local Reconstr.** Patch SSIM	**NME** %	**FR@0.1** % (#)
(R)	ResNet-18				✓	-	-	-	5.64	4.64 (32)
(HG)	Heatmap HG				✓	-	-	-	5.48	4.21 (29)
(A)	Adv. Autoencoder	✓				**12.61**	**0.68**	**0.64**	5.67	4.94 (34)
(A-FT)	Adv. Autoencoder (FT)	✓			✓	25.03	0.57	0.55	4.92	2.47 (17)
(B)	AE + GAN	✓	✓			15.10	0.60	0.58	5.30	3.77 (26)
(B-FT)	AE + GAN (FT)	✓	✓		✓	27.48	0.49	0.50	4.71	2.03 (14)
(C)	AE + GAN + Struct.	✓	✓	✓		15.91	0.62	0.64	4.92	2.61 (18)
(C-FT)	AE + GAN + Struct. (FT)	✓	✓	✓	✓	27.65	0.50	0.53	**4.41**	**1.45 (10)**

Table 1: Results of autoencoder ablation study. Rows (R) and (HG) are benchmark results from fully supervised methods with a comparable ResNet-18 architecture. Rows (A), (B), (C) show the effects of adding loss terms on both global and local reconstruction errors as well as on landmark localization accuracy and failure rate. Rows (A-FT), (B-FT), (C-FT) report results on post-finetuning the autoencoder on the 300-W dataset. NME = Normalized mean error, FR@0.1 = failure rate at 10% NME. All results reported for the full testset of 300-W.

have high visual quality as well as large variability in facial appearance (pose, expression, hair style, accessories).

3. Ablation studies

A critical part of our framework is the first step in which an adversarial autoencoder is trained in an unsupervised fashion on a large dataset of faces, which yields a low-dimensional embedding vector z that encapsulates the face representation.

3.1. Autoencoder losses

The adversarial autoencoder is trained through four loss functions balancing faithful image reconstruction with the generalizability and smoothness of the embedding space needed for the generation of novel faces. A reconstruction loss $\mathcal{L}_{rec}$ penalizes reconstruction errors through a pixel-based $L1$ error. An encoding feature loss $\mathcal{L}_{enc}$ [1] ensures the creation of a smooth and continuous latent space. An adversarial feature loss $\mathcal{L}_{adv}$ pushes the encoder E and generator G to produce reconstructions with high fidelity since training of generative model using only image reconstruction losses typically leads to blurred images. As the predicted landmark locations in our method follow directly from the locations of reconstructed facial elements, our main priority in training the autoencoder lies in the accurate reconstruction of such features, reconstruction accuracy is further enhanced by introducing a structural image loss $\mathcal{L}_{cs}$.

Here, we present results of the framework ablating different loss terms (except for the encoding feature loss $\mathcal{L}_{enc}$) during the training of the autoencoder to study their impact on landmark localization accuracy (see Table 1) using the 300-W dataset. In addition, we report the effects of the optional finetuning step on accuracy, in which the autoencoder is further tuned on the 300-W training dataset. All setups were trained on 128x128px images at a half of the resolution of the setup reported in the paper (see also Figure 2).

As benchmarks, the first two rows of Table 1 also list a standard ResNet-18 predictor of landmark locations (trained on 300-W) as well as a standard heatmap-based system (trained on 300-W). Both approaches offer roughly the same kind of performance on this dataset with a slight advantage for heatmap-based prediction.

If we only add the autoencoder (using $\mathcal{L}_{rec}$, $\mathcal{L}_{enc}$) to our ResNet-architecture, then performance is comparable to that of the standard, non-bottlenecked ResNet-18 architecture, which shows that the 99 dimensions seem to be sufficient to capture the landmark "knowledge" - it is important to note, however, that this landmark knowledge was obtained from *unsupervised* training. Further (supervised) finetuning of the autoencoder on 300-W provides another, significant boost that goes beyond the performance of both supervised benchmark systems. Hence, the finetuning step on the dataset is able to sharpen the implicit landmark representation obtained during the unsupervised step.

Forcing the autoencoder to generate believable images by adding the adversarial loss (using $\mathcal{L}_{rec}$, $\mathcal{L}_{enc}$, $\mathcal{L}_{adv}$) provides a further 7% improvement in NME for standard and finetuned training. Finally, the addition of the structural loss that further enhances small details in the reconstructed faces (using $\mathcal{L}_{rec}$, $\mathcal{L}_{enc}$, $\mathcal{L}_{adv}$, $\mathcal{L}_{cs}$) yields another $\approx$7% improvement. Overall, these results clearly show that losses that tune the face representation to be able to generate more detailed faces will also improve the landmark localization accuracy.

We note that the columns reporting "global" reconstruction errors (as RMSE or SSIM comparisons between the

# Dims	NME %	FR@0.1 % (#)
50†	4.59	**1.02 (07)**
99	**4.41**	1.45 (10)

Table 2: Number of dimension of embedded feature vectors. † Landmark training was instable and required multiple restarts and a reduction of the learning rate.

original and reconstructed images, respectively) and "local" reconstruction error (as SSIM errors evaluated for patches centered on the landmark locations of the original and reconstructed images) yield already good quality for the most "simple" loss setup. For this it is best to look at Figure 2, which shows how the different losses affect the visual quality of the reconstruction. When looking at rows (A), (B), (C), faces gain an increasing amount of high-frequency detail. When adding the GAN loss, these high-frequency details will not aid the reconstruction error at first as the details are "hallucinated" globally all over the face - these details, however, seem to be able to aid the landmark layers in providing a better mapping onto heatmaps and therefore landmark locations. The addition of the SSIM loss does improve the reconstruction error again as the loss forces the high-frequency details to better match with the trained source face images - again, the added details in this case will help landmark localization.

The effect of finetuning on face appearance is interesting to observe as the faces gain immediate detail for all loss setups, yet their overall reconstruction is sometimes more "different" to the source face compared to the non-finetuned version. This is because finetuning unfreezes the weights of the encoder but will train to predict the landmark locations more reliably - hence, the reconstructed faces will favor clear landmark localizability (through well-defined facial feature locations) at the expense of more faithful face reconstruction. Overall, the effect is therefore an increase of the reconstruction error.

As a final note, we observe that training the autoencoder setup on 256x256px provides another jump in performance as the system will learn to reconstruct facial details at an even higher fidelity (see final two rows in Figure 2).

3.2. Encoding length

The latent vector z reported in the main paper has a dimensionality of $d = 99$ which is comparable to other GAN-frameworks [2, 3].

In Table 2, we report the effect of halving this dimensionality to $d = 50$ on landmark localization accuracy. Although yielding a slightly higher NME, the reduced autoencoder obtains a slightly lower FR, which overall means that both embedding dimensionalities result in similar performance levels. An issue with the reduced dimensionality embedding, however, was that the subsequent landmark training was notably less robust, requiring a much more conservative learning rate.

Hence, for the task of landmark localization, the current framework may work with a lower-dimensional embedding space, however, it seems that pulling the implicit information out of the reduced dimensions is a harder task than for a richer embedding.

Further experiments are needed to investigate the effects of increasing the dimensionality as well as providing further constraints on the embedding vector z during the unsupervised training.

3.3. Unsupervised training and few-shot learning

We next take a look at the effects of the unsupervised training step as well as the amount of supervised post-training on 300-W. Table 3 shows again the ResNet-18 and heatmap hourglass baselines and then three different training setups for our full, finetuned system at 128x128px image size.

The first two rows report results of the full architecture without any unsupervised pre-training and hence without any implicit face knowledge. The next rows show results for the full architecture with different amounts of pre-training. Pre-training on the 300-W training dataset results in equal or slightly better performance compared to the baseline architectures showing that the system is able to pick up implicit knowledge already from only 3,200 images. Pre-training on 100,000 images provides a significant, further jump as does pre-training on the full 2,1M image dataset.

Importantly, the error increase in the presence of limited training data (columns labeled 1.5% in Table 3) with just 50 images showcase the power of the pre-trained representation: whereas ResNet-18 increases around 54% in NMW from 100% to 1.5% training set size, our pre-trained architectures only reduce 47%, 34%, and 29% respectively owing to the more robust generalization from the latent representation.

4. Few-shot learning on different datasets

Figures 3,4,5 show results for few-shot learning on the three different datasets (300-W, AFLW, WFLW) reported in the main paper. The first column has the *entire* training set (50, 10, or 1 labeled image(s)), and the second column shows predicted landmarks on nine or three images from the different testsets contained in the datasets. In all figures, training with even just one image produces reasonable localization results and a clear improvement in prediction accuracy can be traced as a few more images are added.

In Figure 5, the failure cases are most visible (see, for example, the top results for training with one image on the Blur testset). It should be noted that this is by far the

		Unlabeled training data			Labeled training data			
Model	**Num. param.**	**Pre-train dataset(s)**	**Num. of images**	**External images**	**100% (3,189)** NME	FR@0.1	**1.5% (50)** NME	FR@0.1
ResNet-18	11M	None	0	no	5.64	4.64 (32)	8.70	22.21 (153)
Heatmap HG	22M	None	0	no	5.48	4.21 (29)	10.13	39.33 (271)
C-FT	23M	300-W	3,189	no	5.40	4.79 (33)	7.95	15.82 (109)
C-FT	23M	VGG + AN	100k	yes	4.73	1.74 (12)	6.34	9.29 (064)
C-FT	23M	VGG + AN	2.1M	yes	**4.41**	**1.45 (10)**	**5.71**	**4.35 (030)**

Table 3: Effect of unsupervised pre-training when trained with full and reduced labeled training data on 300-W.

most challenging dataset as it contains variability in face appearance (due to illumination, occlusion, and make-up) that is not fully present in the unsupervised datasets we used (cf. the randomly-generated faces in Figure 1). As a few more labeled faces are added, however, performance begins to quickly improve even in the presence of such severe changes.

5. Additional result visualizations

Figures 6,7,8 show additional, non-curated visualizations of the full system on images from the six test subsets of WFLW.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 2

[2] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 3

[3] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 3

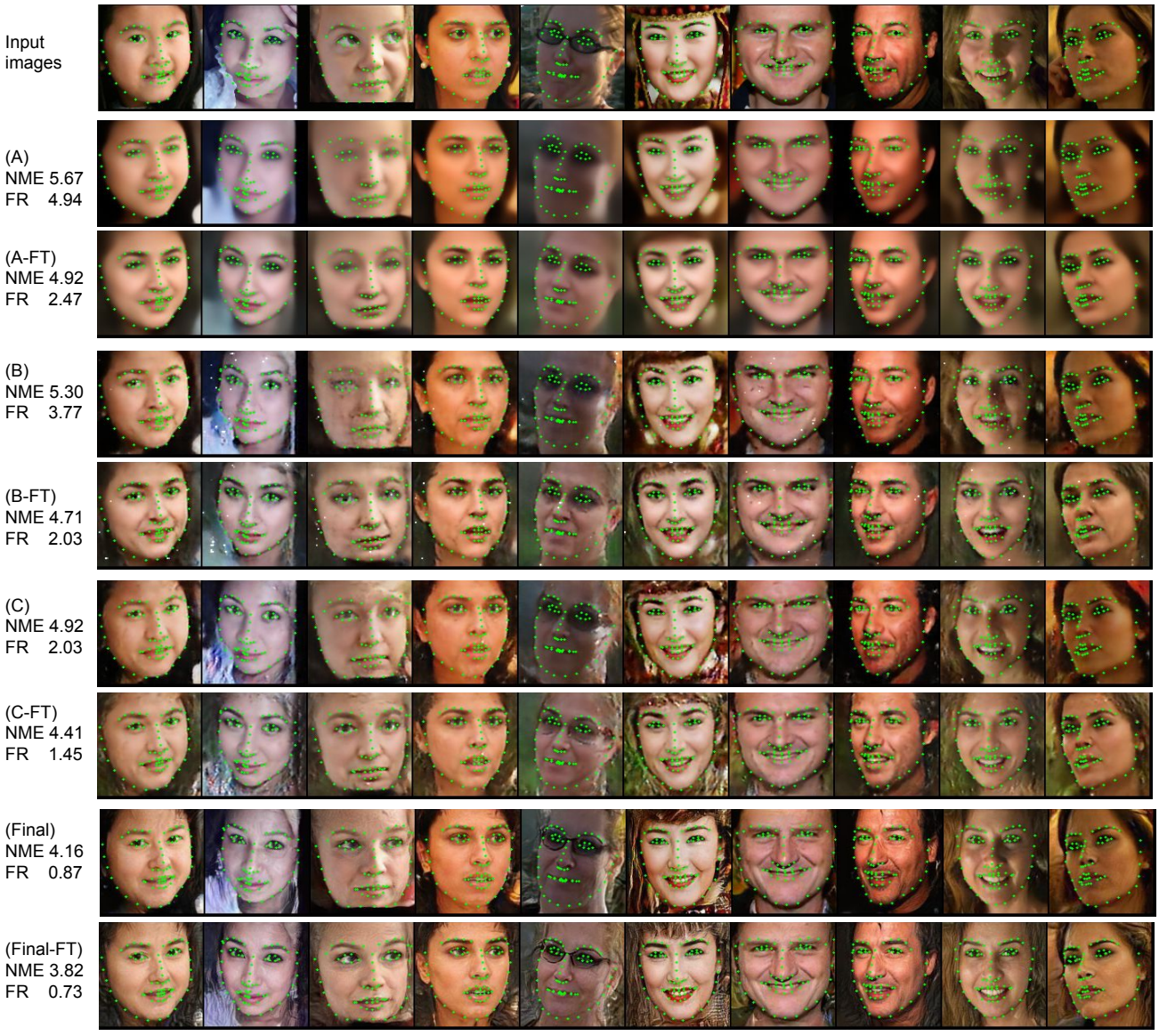

Figure 2: Example reconstructions corresponding to Tab. 1. (A)-(C) are trained for 30 epoches on 128×128 images. 'Final' denotes the fully trained model 256×256 that was used for the experiments in Sec. 4 of the paper.

Figure 3: Few-shot learning on 300-W

Figure 4: Few-shot learning on AFLW

Figure 5: Few-shot learning on WFLW

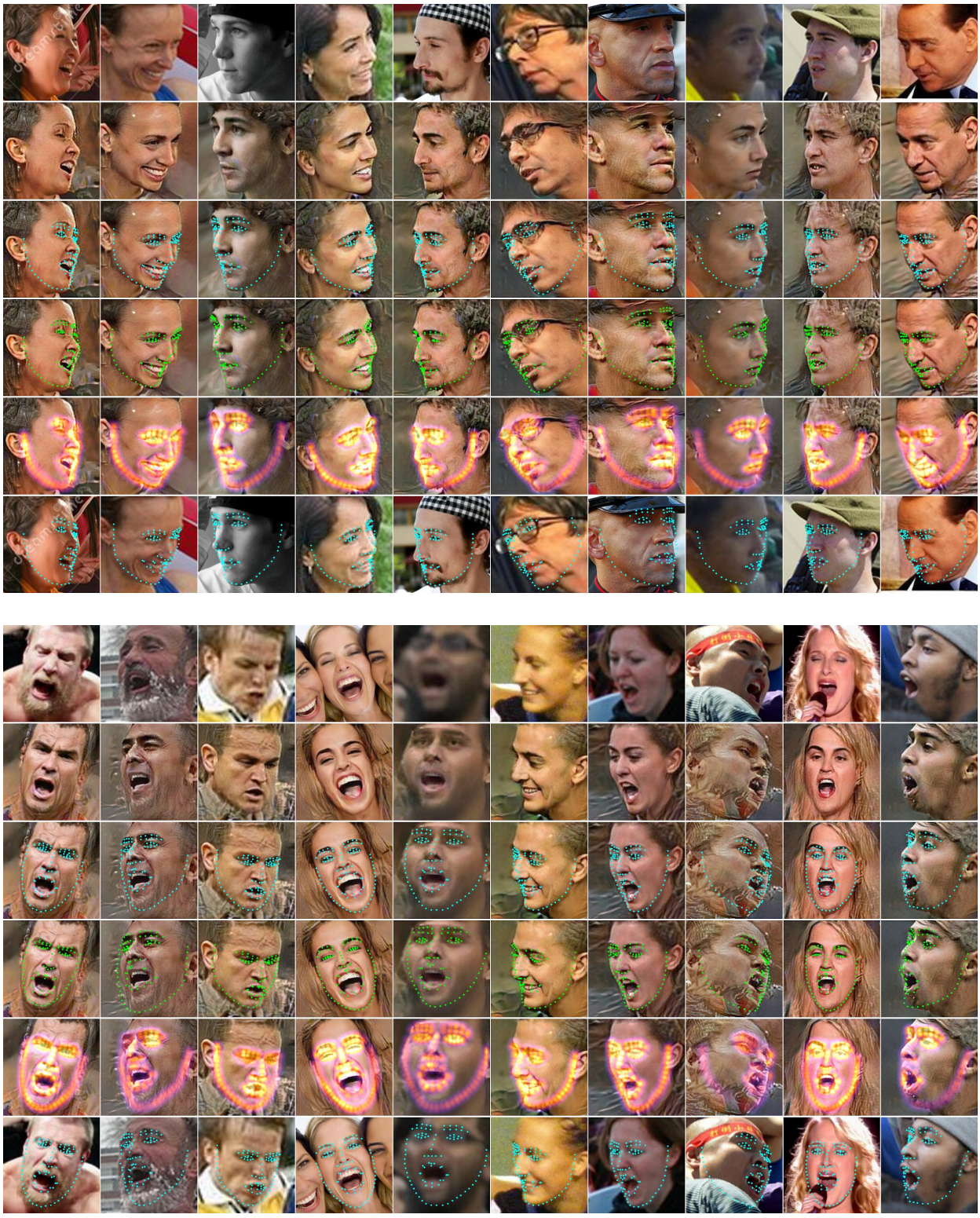

Figure 6: 3FabRec results on WFLW Pose and Expression: Two blocks of rows show (1) original, (2) reconstruction, (3) reconstruction with predicted landmarks, (4) reconstruction with ground-truth landmarks, (5) reconstruction with predicted landmark heatmaps, and (6) original with predicted landmarks, respectively.

Figure 7: 3FabRec results on WFLW Illumination and Make-Up: Two blocks of rows show (1) original, (2) reconstruction, (3) reconstruction with predicted landmarks, (4) reconstruction with ground-truth landmarks, (5) reconstruction with predicted landmark heatmaps, and (6) original with predicted landmarks, respectively.

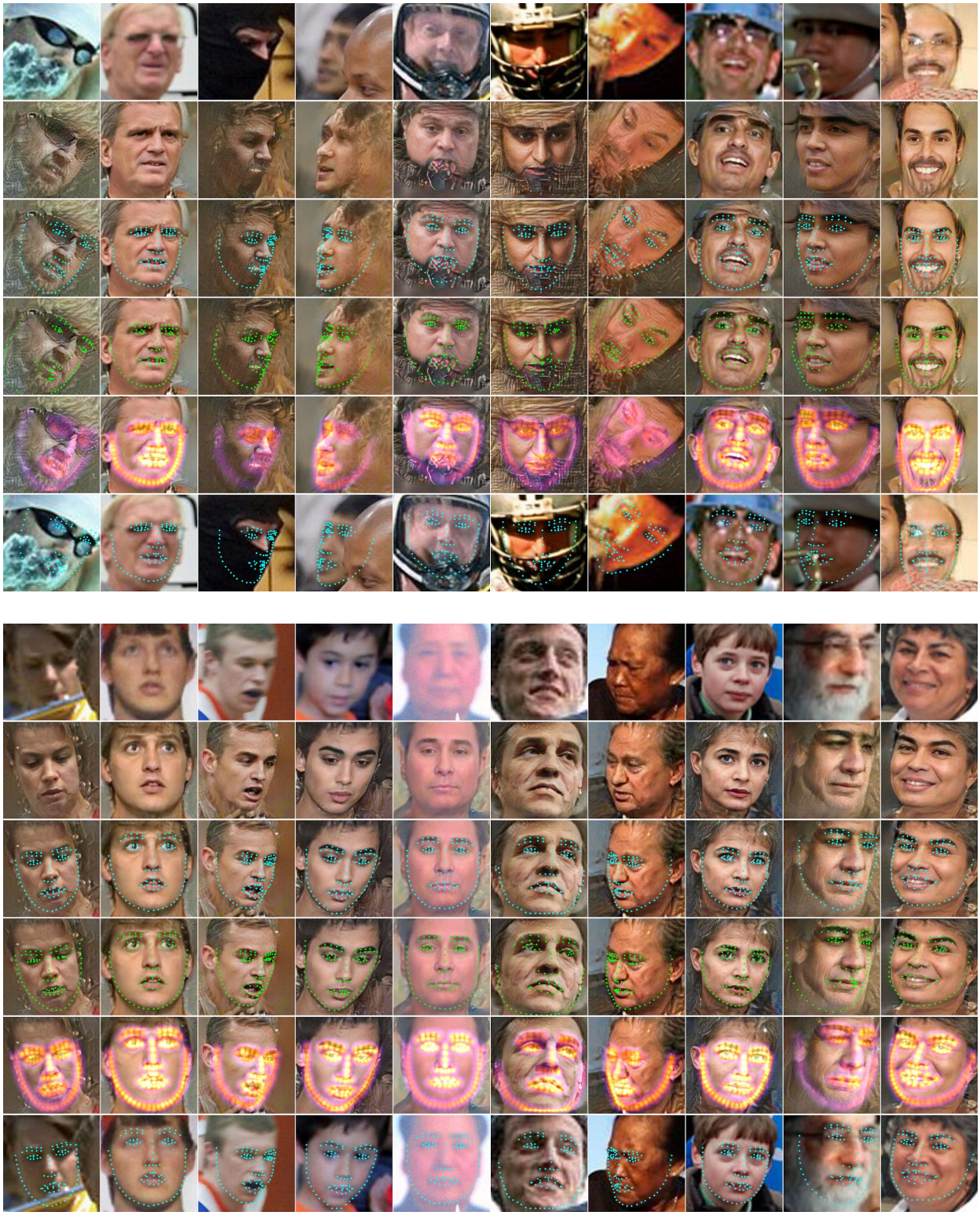

Figure 8: 3FabRec results on WFLW Occlusion and Blur: Two blocks of rows show (1) original, (2) reconstruction, (3) reconstruction with predicted landmarks, (4) reconstruction with ground-truth landmarks, (5) reconstruction with predicted landmark heatmaps, and (6) original with predicted landmarks, respectively.