Supplementary material: C3AE: Exploring the Limits of Compact Model for Age Estimation

Chao Zhang[1,2], Shuaicheng Liu[1,2], Xun Xu[3], Ce Zhu[1,*]
University of Electronic Science and Technology of China[1] Megvii Technology[2]
National University of Singapore[3]
galoiszhang@gmail.com, liushuaicheng@megvii.com, eczhu@uestc.edu.cn, elexuxu@nus.edu.sg

1. Additional Results of Ablation Study I

1.1. Plain Model

In this section, some detailed results and explanations are included to compare with our plain model. In order to validate our claims that for small size image and model, standard convolution performs better than depth-wise convolution, several variants on MobileNet-V2 and ShuffleNet-V2 are compared with the plain model of C3AE. In addition, we also compare with the competitive SSR. The curve of train/validation losses on MobileNet-V2, ShuffleNet-V2, SSR are given in Fig. 1, Fig. 2, Fig. 3 (Some partial results are shown in Fig. 4 (main manuscript)). On one hand, for any finetuning of factors in Tab. 2 (main manuscript), the plain model of C3AE consistently achieves better result. On the other hand, considering the gap between train and validation loss, our plain model is the lowest. We could argue that the plain model of C3AE is more robust on the generalization.

From Fig. 3, our plain model performs better than SSR, both with standard convolution. In the experimental implementation, SSR adopts the full model instead of the plain model. This comparison further shows the superiority of our plain model.

In general, Tab. 2 (main manuscript), Fig. 1, Fig. 2 and Fig. 3, with three large datasets IMDB (460,000 images), WIKI (62,000 images) and Morph II(55,000 images), supports the effectiveness of our claims.

1.2. Residual Module and SE Module

To support two viewpoints in Section 3.5 (main manuscript), the residual module and Squeeze-and-Excitation (SE) module are implemented with ablation study, as shown in Fig. 4. For three datasets, SE module plays the positive role, while residual module works negatively towards the final result. The results in Tab. 3 (main manuscript) and Fig. 4 demonstrate our assumptions.

*Corresponding author

2. Additional Results of Ablation Study II

In Fig. 5 (main manuscript), we plot the comparative results on with/without cascade module and with/without context module. Here the exact values are given in Tab. 1 and Tab. 2. In fact, the cost of parameters and memory on cascade and context module are negligible.

Table 1. With/without cascade and context module.

Methods	MAE	Memory	Parameters
w/o-cascade+SE	2.98	0.23MB	39.4K
cascade+SE	2.92	0.24MB	39.5K
cascade+SE+context	2.75	0.25MB	39.7K

Table 2. Different λ and context conditions.

Methods	$\lambda = 5e$-5	$\lambda = 5e$-4	$\lambda = 0.005$	$\lambda = 0.05$	$\lambda = 0.5$	$\lambda = 5$
w/o context	2.94	2.93	2.92	2.95	2.95	2.97
context	2.77	2.76	2.75	2.79	2.80	2.84

To measure the sensitivities of C3AE, we finetune the hyperparameters α in Eq.6 (main manuscript). As shown in Tab. 3, five different parameters $\alpha = 5, 8, 10, 12, 15$ on the full model of C3AE are tested. Their results remain stable, and demonstrate the robustness of C3AE.

Table 3. Different α and context conditions.

Methods	$\alpha = 5$	$\alpha = 8$	$\alpha = 10$	$\alpha = 12$	$\alpha = 15$
C3AE full model	2.94	2.93	2.92	2.95	2.95

Some additional examples are given in Fig. 5, i.e., the supplement of Fig. 6 (main manuscript). From all these examples, two or three adjacent elements are nonzero. That is to say, two points representation gives ideal constraints on the age distribution, and rule out some unwanted or negative representations like $50 = 0.5 \times 0 + 0.5 \times 100 = 0.2 \times 25 + 0.2 \times 50 + 0.2 \times 75 + 0.2 \times 100$. The cascade module plays an important role on controlling the diversified combinations of age representation. In addition, context based result is superior to the vanilla result.

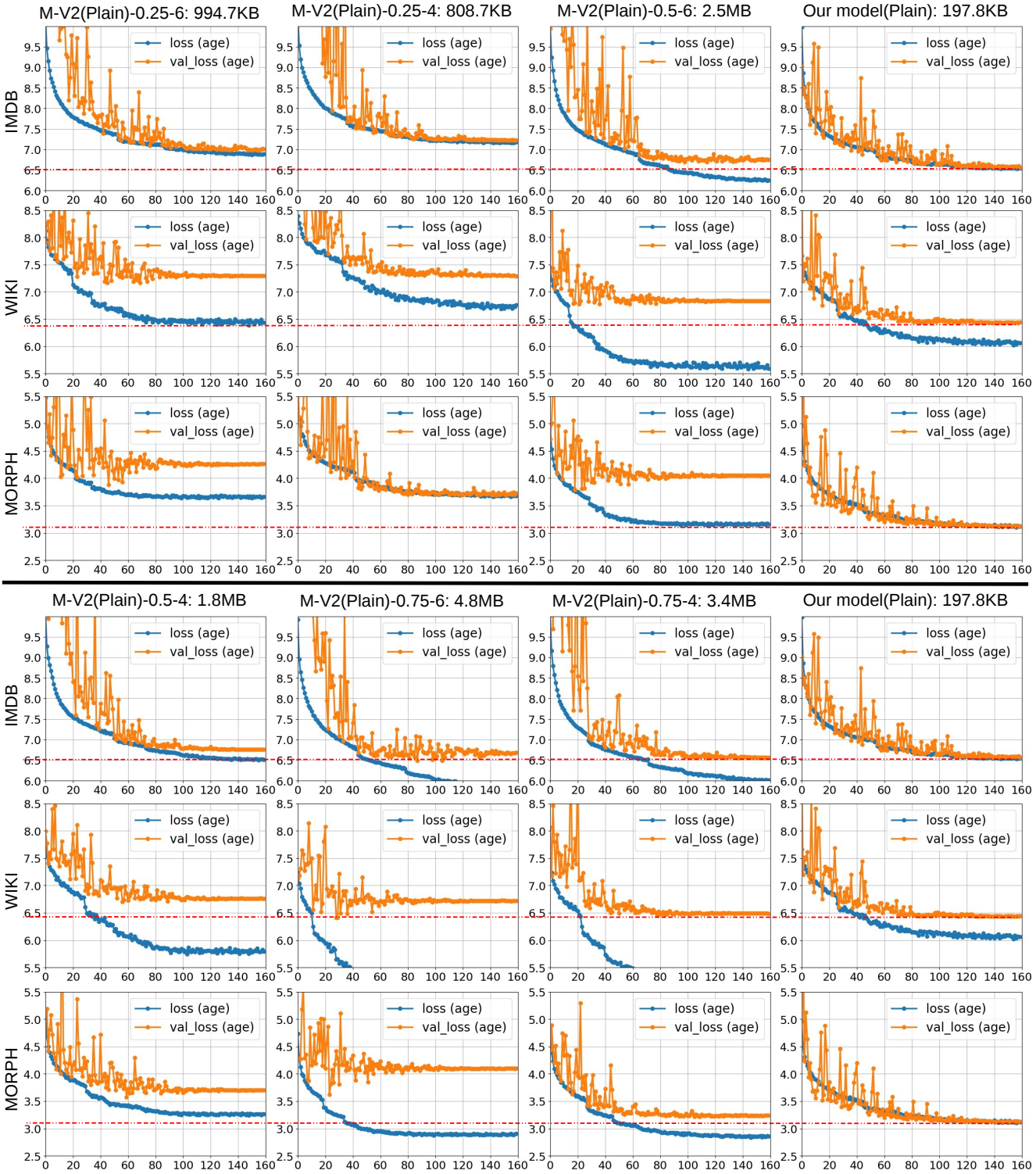

Figure 1. Comparison with MobileNet-V2. The curves in blue and orange mean the train and validation loss, respectively. The red baseline is the validation loss of our plain model. We consider two aspects: the validation loss and the gap between train and validation loss. Our plain model consistently performs better than MobileNet-V2. (Best viewed in color and magnifier.)

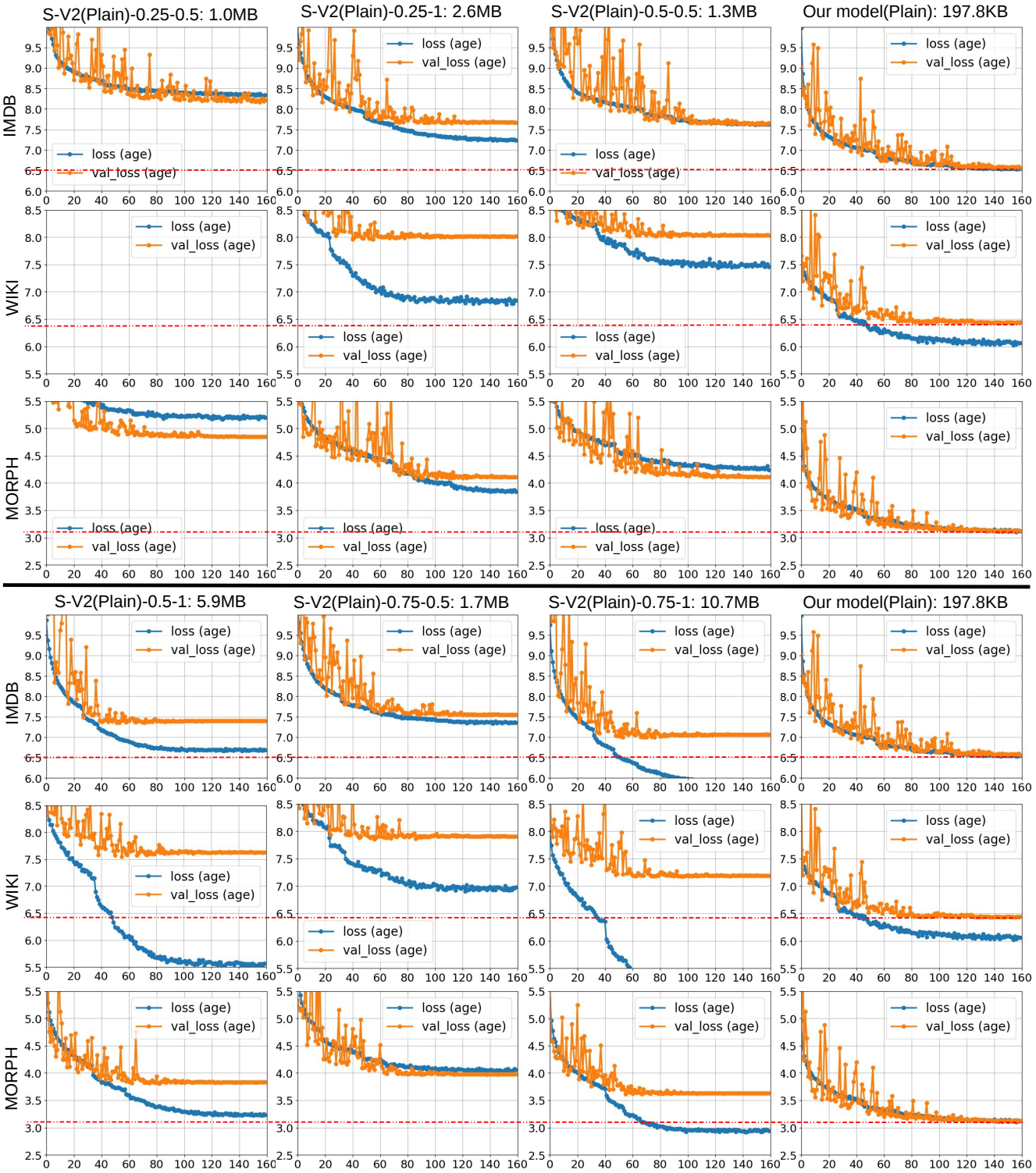

Figure 2. Comparison with ShuffleNet-V2. The curves in blue and orange mean the train and validation loss, respectively. The red baseline is the validation loss of our plain model. We consider two aspects: the validation loss and the gap between train and validation loss. Our plain model consistently performs better than ShuffleNet-V2. (Best viewed in color and magnifier.)

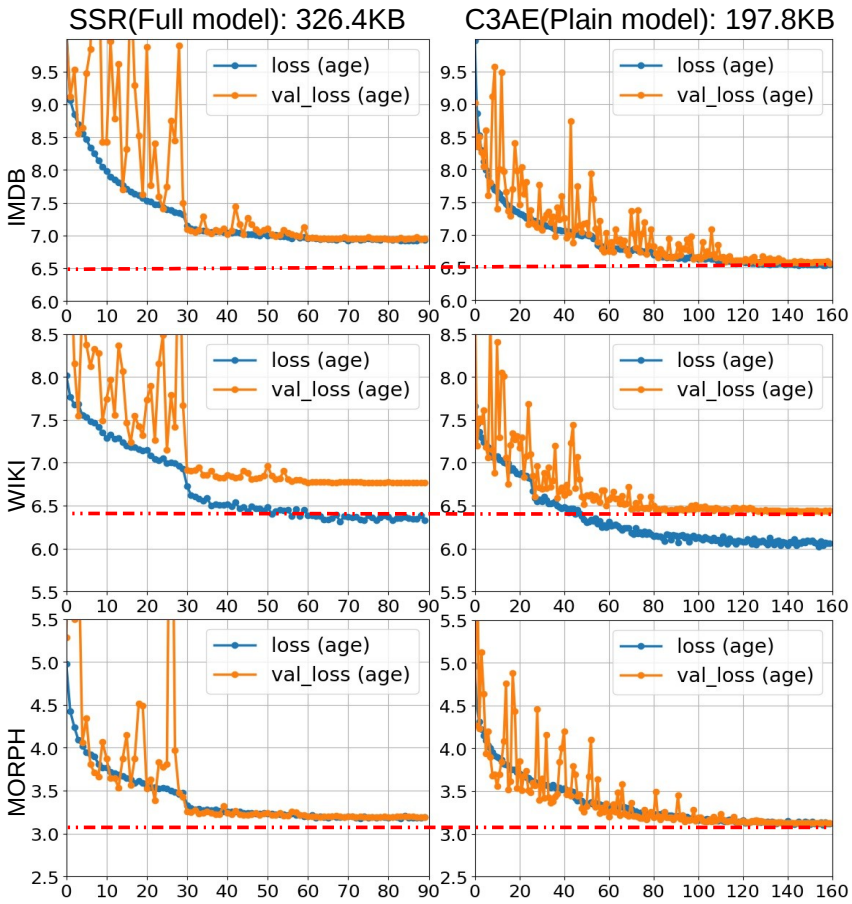

Figure 3. Comparison with SSR. The curves in blue and orange mean the train and validation loss, respectively. The red baseline is the validation loss of our plain model. We consider two aspects: the validation loss and the gap between train and validation loss. Our plain model even gets better result than SSR. (Best viewed in color and magnifier.)

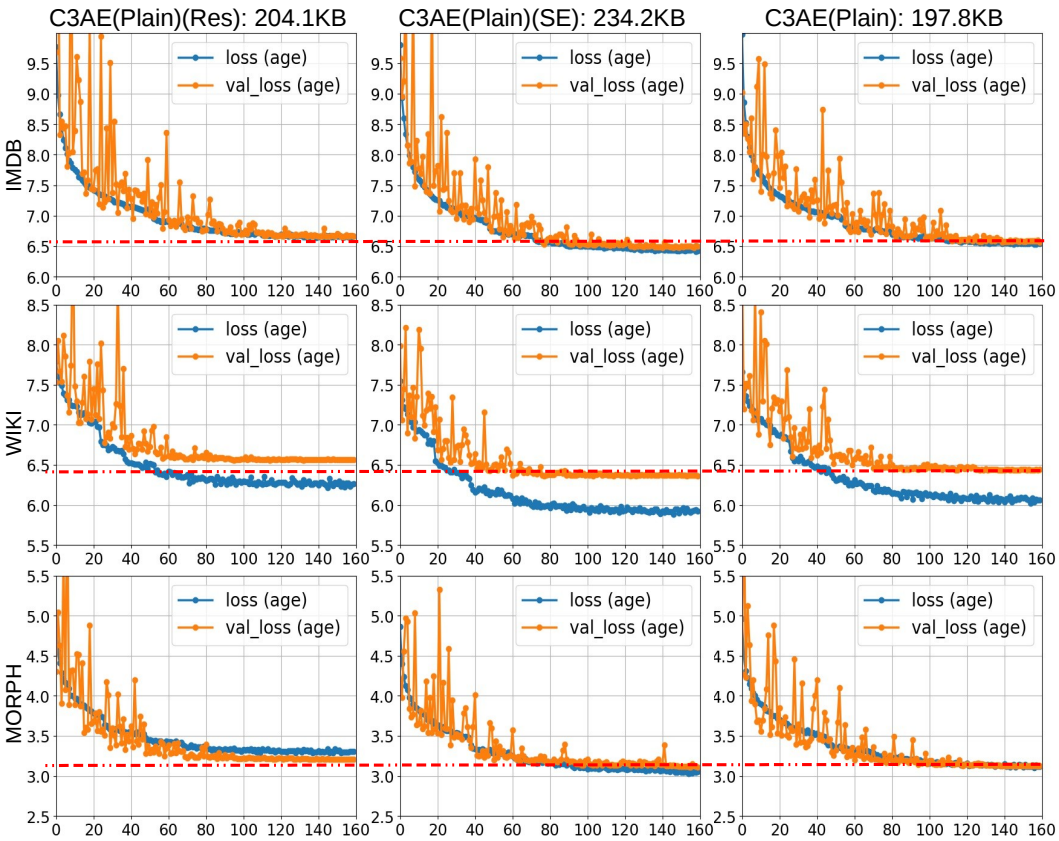

Figure 4. With/without residual module and SE module. The curves in blue and orange mean the train and validation loss, respectively. The red baseline is the validation loss of our plain model. We consider two aspects: the validation loss and the gap between train and validation loss. Compared with the plain model, the residual module and SE module get negative and positive results, respectively. (Best viewed in color and magnifier.)

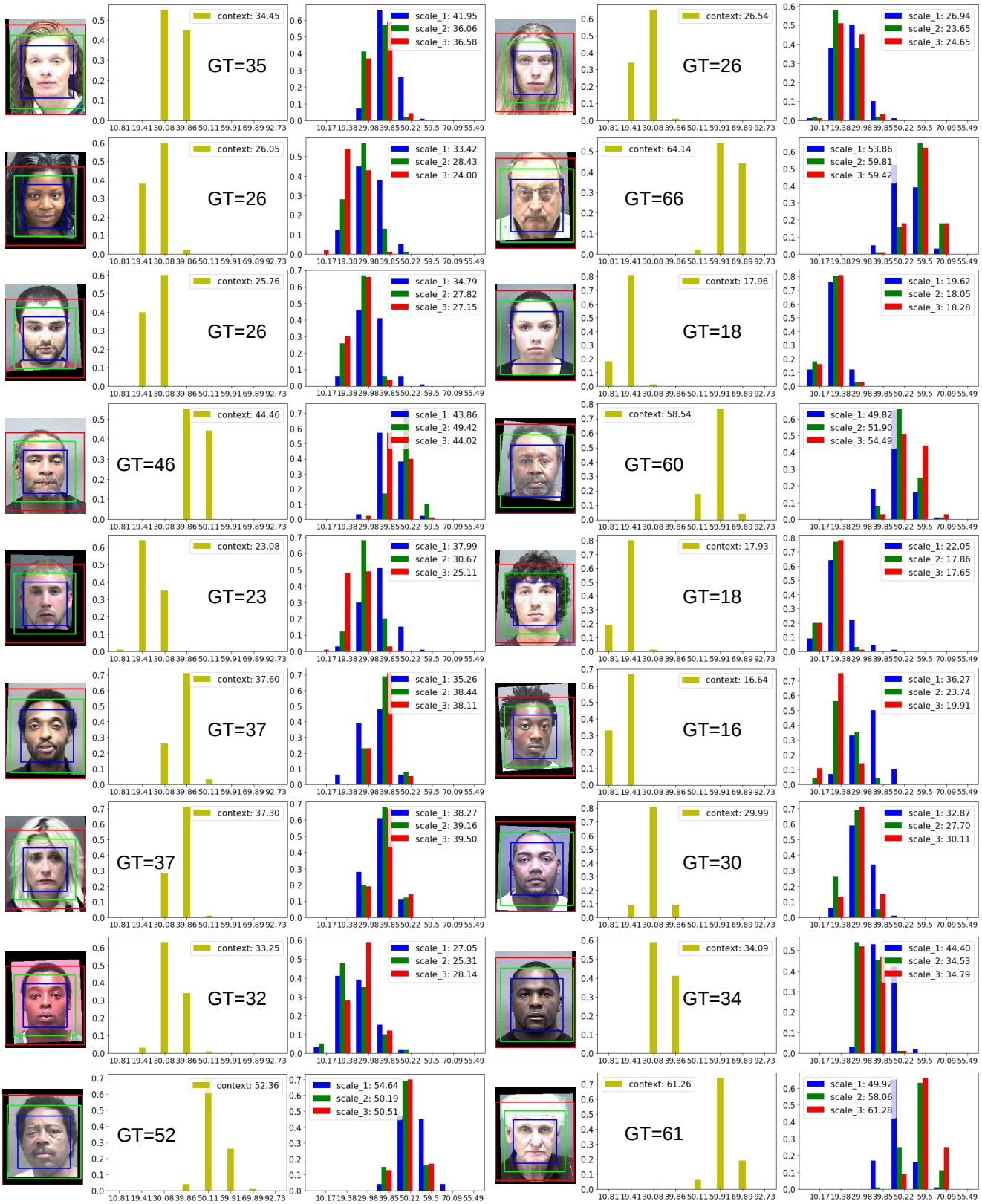

Figure 5. Some additional examples for Fig. 6 (main manuscript) and Section 4.3.2 (main manuscript). For each facial image, two distributions: context/non-context based vector, are given at the second and third column. The former is with yellow color and the latter are with red/green/blue color corresponding to three different contexts (or three resolutions corresponding to three colored bounding boxes on the facial image). In specific, the latter only uses single resolution image while the former inputs three context images. The former always performs better than the latter. In other words, context module works well. The fact that there are two or three adjacent elements are nonzero supports cascade module. (Best viewed in color and magnifier.)