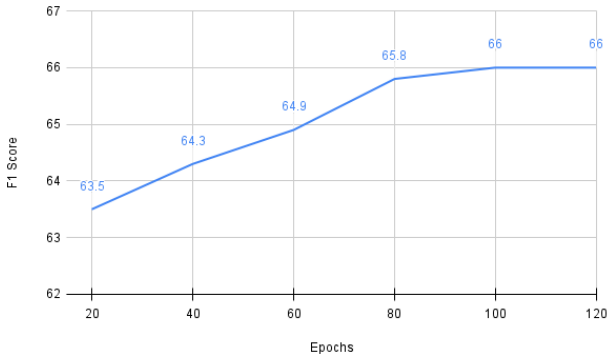
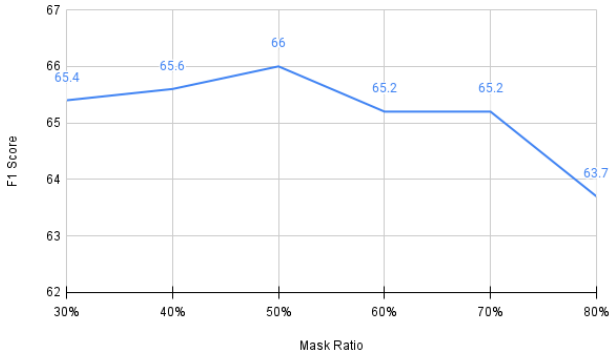
Multimodal Channel-Mixing: Channel and Spatial Masked AutoEncoder on Facial Action Unit Detection Supplementary Material

1. Mask ratio and Training epoch in pre-training



(a) Fine-tune performance on various pre-training epochs, where a noticeable improvement along epochs.



(b) Fine-tune performance on various mask ratios of the pre-training model.

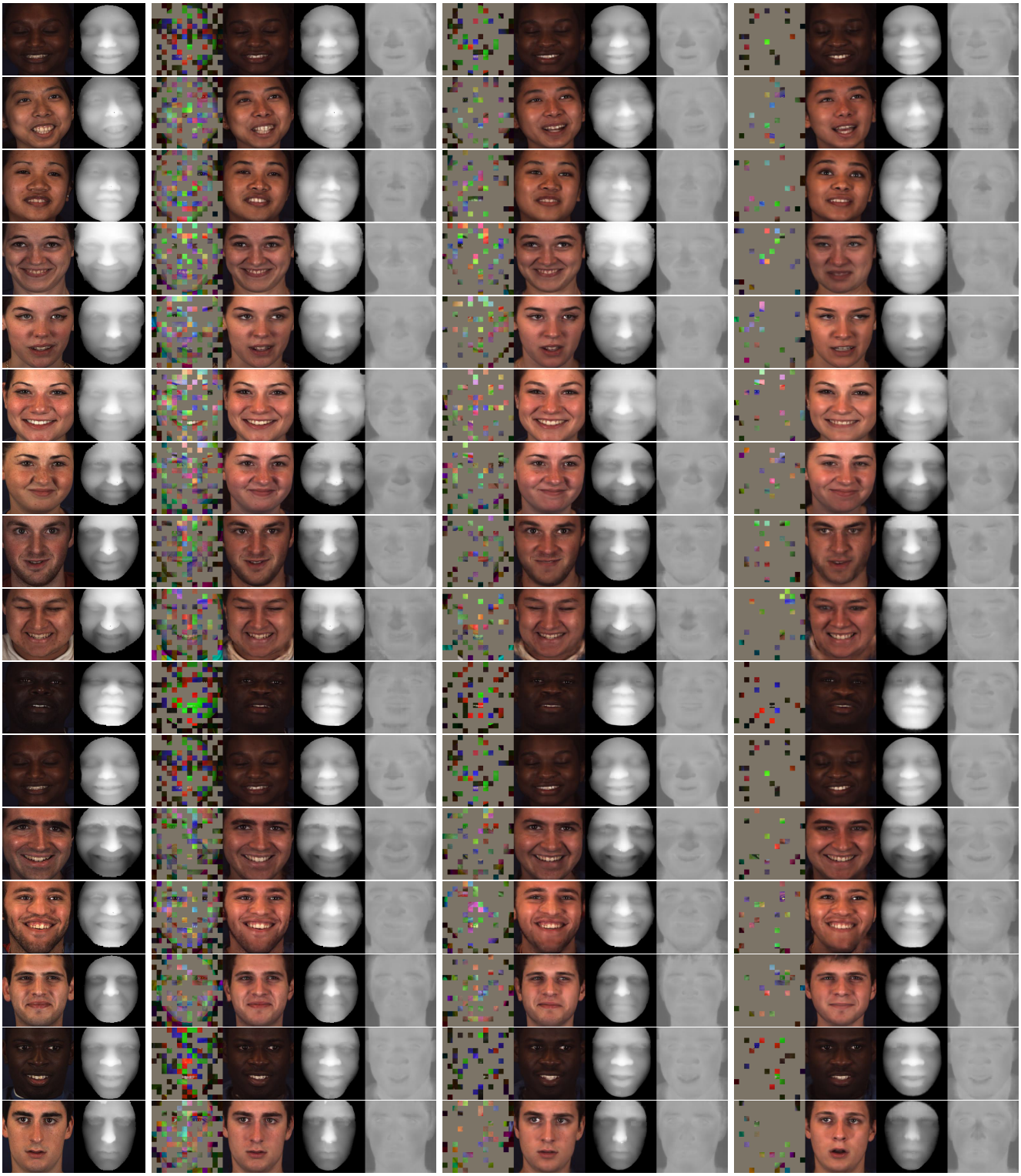
Figure 1. (a) F1 scores are reported by the model trained on BP4D+ with 50% mask. (b) F1 scores in fusion are reported by the model trained on BP4D+ with epoch-100.

First, the ablation study of pre-training epoche is based on a 50% mask, where the result of RGB modality is shown in Figure 1a in terms of average F1 scores. Performance steadily improved during the training process and reached its optimum at the 100 epoch. Our mask ratio experiments thus far are based on the pre-training model on epoch-100.

Then we evaluate the masking ratio influence on downstream AU detection performance, see in Figure 1b. Performance is improved as the mask ratio is increased, while the optimal ratio is 50%. When the ratio is greater than 50%, the f1 score is obviously decreased, because much more detailed information is missed. Our following reconstruction visualization is based on the model training on BP4D+ with 100-epoch and 50% mask ratio.

2. More Visualization Examples

More visualization examples on BP4D and DISFA are shown in Figure 2 and Figure 3.



(a) origin (b) mask: 50% (c) mask: 75% (d) mask: 90%

Figure 2. Visualization of the reconstructed images on BP4D with different mask ratios is performed using a reconstruction model trained on BP4D+. Images in (b, c, and d) represent the masked channel-mixing image, reconstructed RGB, reconstructed Depth, and reconstructed Thermal, respectively.

(a) mask: 25% (b) mask: 50% (c) mask: 75%

Figure 3. Visualization of the reconstructed images on DISFA with different mask ratios is performed using a reconstruction model trained on BP4D+. Images in each cell represent the original RGB, masked image, reconstructed RGB, reconstructed Depth, and reconstructed Thermal, respectively.