Supplementary Material for Latent-OFER: Detect, Mask, and Reconstruct with Latent Vectors for Occluded Facial Expression Recognition

Isack Lee, Eungi Lee, Seok Bong Yoo*
Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea
{sackda24, 181061, sbyoo}@jnu.ac.kr

We present the effect of semantic consistency loss and additional results in this supplementary materials. We provide the effects of expression-relevant feature extractor; comparison of results on real-world occluded facial images; detailed comparative analysis about real-world occlusion facial expression recognition (FER) result.

1. Effect of semantic consistency loss

To examine the effect of semantic consistency loss, we select the RAF-DB [3] images and train hybrid reconstruction networks by varying the setting of the trade-off parameter λ_{sc}. Figure 1 depicts a visual comparison of the reconstruction results obtained using different values of λ_{sc}. The figure highlights the variations in image quality, and semantic consistency as the value of λ_{sc} changes. The λ_{sc}= 0 indicates that semantic consistency loss is not used. If λ_{sc} is set to 10, indicating an over-weighting of the semantic consistency loss, the resulting reconstructions tend to exhibit semantically detrimental effects. In contrast, other values of λ_{sc} (i.e., 0, 0.1, and 1) produce visually pleasing reconstruction results. Table 1 presents a comparison of the effect of λ_{sc} value on the accuracy of FER under the same conditions, except for the parameter λ_{sc}. To ensure the highest accuracy in recognizing facial expressions, we selected a value of λ_{sc}=1.

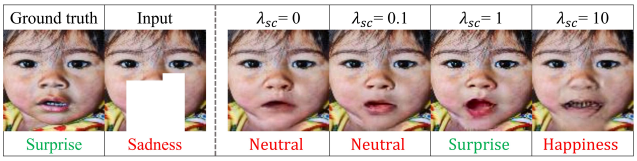

Figure 1. Qualitative comparison of reconstruction results obtained by changing the λ_{sc} setting on Grad-occlusion masked RAF-DB.

*Corresponding author

λ_{sc}	0	0.1	1	10
FER accuracy	78.9	79.2	80.1	77.3

Table 1. FER accuracy (%) for reconstruction results of Grad-occlusion masked RAF-DB for hybrid reconstruction networks trained with different λ_{sc} values, respectively.

2. Additional results

2.1. Effect of expression-relevant feature extractor

Our proposed expression-relevant feature extractor utilizes only the ViT-latent vectors that are relevant to facial expressions from the entire ViT-latent space. The proposed approach is expected to improve FER performance by avoiding the use of irrelevant information to facial expression, such as hairstyle and background. Additionally, the proposed method allows for the exclusion of information from incompletely reconstructed regions, effectively preventing a degradation in performance.

To demonstrate the effectiveness of expression-relevant feature extractor, we combine all recognition models with a deocclusive autoencoder and compare their performance. As shown in Table 2, our proposed model outperforms the others, even when recognizing facial expressions on the same reconstructed images. This is because expression-relevant feature extractor excludes non-expressive information, which is detrimental to learning. Moreover, the combination of expression-relevant latent vectors and CNN features achieves superior accuracy by learning cooperatively. The results indicate the performance improvement is not solely due to the deocclusive autoencoder, but rather the cooperative learning of expression-relevant feature extractor.

2.2. Comparison of results on real-world occluded facial images.

Figure 2 shows the real-world occluded and reconstructed images, alongside their corresponding predictions. The images are presented to compare the adequacy of the reconstruction in terms of FER. The hybrid reconstruction network utilizes a self-assembly layer to remove occlusions

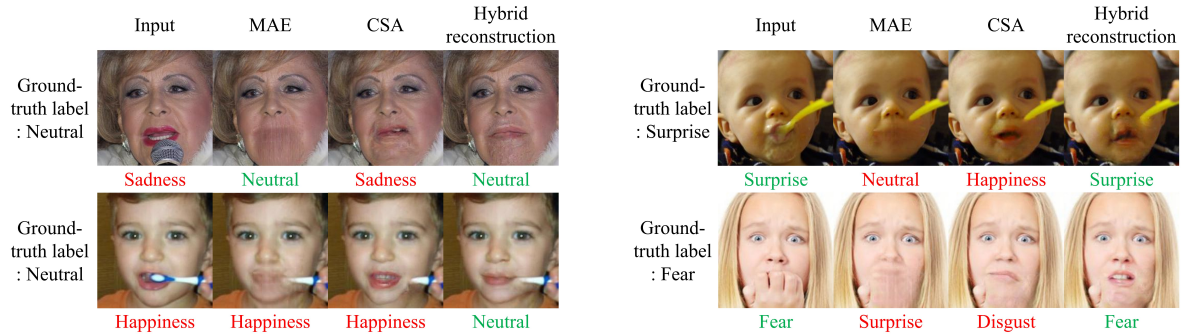

Figure 2. Comparison of results on real-world occlusion in AffectNet [6]. Predictions of FER according to image reconstruction. The leftmost side of each image group indicates the ground truth label. Labels highlighted in green indicate matching the correct expression, whereas red indicates a misprediction.

	Deocclusive autoencoder + DACL [1]	Deocclusive autoencoder + RUL [8]	Deocclusive autoencoder + DAN [7]	Deocclusive autoencoder + EAC [9]	Latent -OFER
Syn-AffectNet	50.6	54.1	51.5	51.9	56.1
Syn-RAFDB	72.4	77.6	78.5	79.1	80.1
Syn-KDEF	83.4	85.2	84.6	85.6	86.7

Table 2. Accuracy (%) comparison images deocluded with a deocclusive autoencoder on the Syn-FER dataset.

and generate realistic images, while also incorporating semantic consistency loss to ensure proper facial expressions. The results shown in Figure 2 indicate that occlusions can cause inaccuracies in predicting facial expressions. The reconstruction results of MAE [2] tend to be blurry. The blurred region do not capture high-frequency details, making it difficult for FER systems to accurately distinguish subtle changes in facial expressions, thereby decreasing the recognition accuracy. Although the CSA [5] method is capable of generating believable facial images, it does not take into account the facial expressions, leading to distortion in the expression. Therefore, we conclude that the hybrid reconstruction network is successful in reconstructing occluded regions of the facial image, resulting in improved FER.

2.3. Detailed comparative analysis of results in FED-RO

Our proposed model shows robustness in real-world scenarios, as we have validated using the FED-RO [4] dataset. We conducted a thorough analysis of the prediction result using a normalized confusion matrix. We present the FER results of various occluded FER models in Figure 3. Our proposed model outperforms other models for most expressions, especially neutral and anger. However, we observed relatively poor performance for fear and disgust compared to other models. In particular, fear is misrecognized as surprise 40% of the time, and disgust is frequently misrecognized as sadness and anger.

Figure 4 provides insight into why fear expression is frequently misidentified as surprise in the model and why recognition accuracy for disgust is low. We attribute this to the presence of noisy labels in the FED-RO dataset, particularly for fear and disgust. Figure 3 also shows that predictions of Latent-OFER are often more reliable in such cases. Thus, we conclude that the lower recognition performance of Latent-OFER for fear and disgust is not a limitation of the model but rather the result of the presence of noisy labeled data.

References

[1] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2402–2411, 2021.

[2] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[3] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.

[4] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Occlusion aware facial expression recognition using cnn with attention mechanism. *IEEE Transactions on Image Processing*, 28(5):2439–2450, 2018.

[5] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4170–4179, 2019.

[6] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

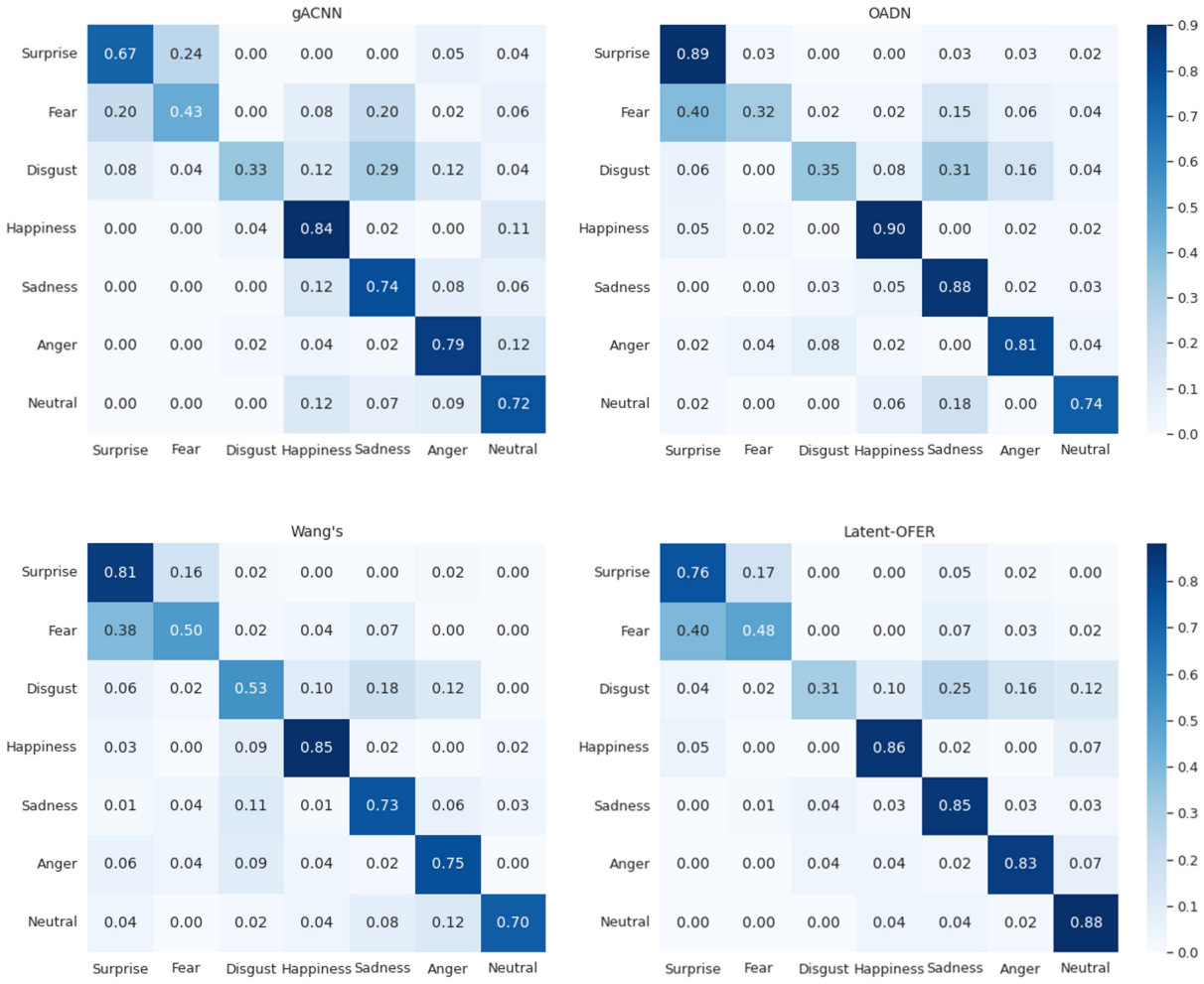

Figure 3. Accuracy (%) comparison for different models with nomalized confusion matrix. (a) gACNN, (b) OADN, (c) Wang's, and (d) Latent-OFER FER results. The vertical axis represents ground truth labels and the horizontal axis represents prediction labels.

[7] Zhengyao Wen, Wenzhong Lin, Tao Wang, and Ge Xu. Distract your attention: Multi-head cross attention network for facial expression recognition. *arXiv preprint arXiv:2109.07270*, 2021.

[8] Yuhang Zhang, Chengrui Wang, and Weihong Deng. Relative uncertainty learning for facial expression recognition. *Advances in Neural Information Processing Systems*, 34:17616–17627, 2021.

[9] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 418–434. Springer, 2022.

Figure 4. Example of noise labels for FED-RO. Labels highlighted in green indicate matching the ground-truth label, whereas red indicates a our model's prediction.