OpticalDR: A Deep Optical Imaging Model for Privacy-Protective Depression Recognition - Supplemental Material

A. Training Setup

A.1. Lens Parameters Setup

In our simulation, the wavefront is modeled with a pixel size of 3.69 μm and a resolution of 1024 × 1024. We use the first 15 terms in Noll notation to represent the Zernike coefficients, which model the surface profile. The refractive indices are defined as 1.488, 1.493, and 1.499 to simulate the lens's refraction of red, green, and blue light at wavelengths of 640 nm, 550 nm, and 460 nm, respectively. The depth is randomly sampled from a range from 0.33 to 2 meters to account for potential variations in distance between individuals and the camera in real-world scenarios. The distance between the lens and sensors z is set at 35.5 mm.

A.2. Initialization, Training and Fine-Tuning Setup

The training of OpticalDR consists of four steps. The training details of each steps are list as follows:

Step 1. We utilize the CelebA dataset [2] to train the lens and SANet, allowing the lens to generate privacy-preserving images with L_i. The lens is initialized with the fourth Zernike coefficient set to -51 and is trained using the Adadelta optimizer with a learning rate of 1. Concurrently, SANet is trained with the Adam optimizer employing a learning rate of 0.01. We save all parameters when the validation loss reaches a minimum. The SANet structure used in this stage is a compact version of ResNet, specifically ResNet10, with bottleneck blocks replaced by SA modules.

Step 2. The CK+ dataset [3] is employed to train the lens and a SANet for acquiring emotional information. We use pretrained parameters of the lens and SANet from *Step 1* as the initial lens and the initial emotion recognition model. The learning rate for the lens is set to 0.01, while for SANet, it is set to 0.0001 during this stage. Parameters are saved when the validation loss L_e reaches a minimum.

Step 3. The AVEC 2014 dataset [7] is utilized for acquire depression-related features optimizing L_d. Videos from AVEC 2014 are extracted frame-by-frame and employed for training purposes. During the validation process, videos are sampled with a frame interval of 10, and the average output from all frames within a single video sample is used as the result for that sample. Human face alignment is performed using the Dlib toolkit. During alignment, we ensure that the centers between the eyes are aligned, and the vertical distance between the eyes and the mouth is set to be 1/3 of the image height. In the training phase, we use the pretrained lens and SANet parameters obtained from *Step 2* as the initial model weights containing emotion information. The learning rate for the lens is set to 0.01, and for SANet, it is set to 0.0001 in this stage. All parameters are saved when the validation loss L_d reaches a minimum.

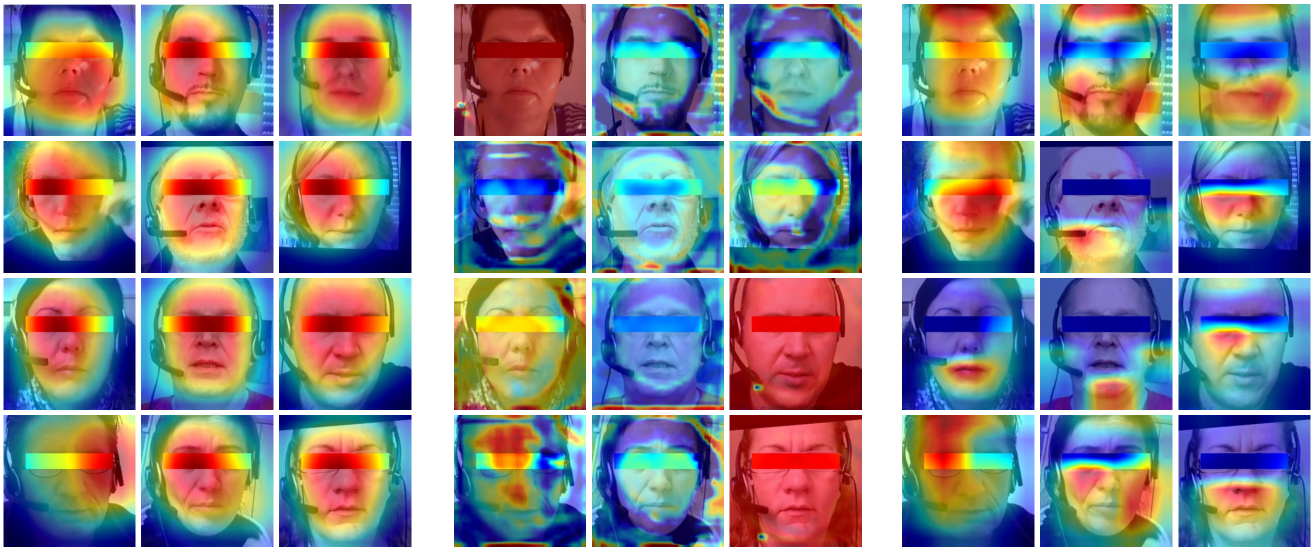
Step 4. For fine-tuning with the final fusion layer, we utilize the AVEC 2014 dataset and employ the depression self-evaluation score of each sample provided by the dataset as the label. In this step, we employ the fusion model with 2 MulT [6] layers, each having 4 attention heads. The lens uses the weight from *Step 3*, and the emotion SANet and depression SANet use weights from *Step 2* and *Step 3*, respectively. Then the parameters of lens and SANets are frozen during training in this step. Training of the fusion layer is performed using Adam as the optimizer with a learning rate of 0.0001 for optimizing L_s. During validation and testing, we assess performance using MAE and RMSE as the evaluation criteria on AVEC 2013 and AVEC 2014, allowing for comparisons with other approaches. The calculation method for validation and testing involves averaging the output from frames sampled at intervals of 10.

A.3. Comparison with Facial Features Disentanglement

Since we progressively train the OpticalDR, during which we involve the disentanglement of facial features, emotion, and depression features, we assessed the performance of TDGAN [8], which employs the disentanglement of identity and emotion features for facial expression recognition. The results are presented in Tab. 1. TDGAN demonstrates competitive performance, particularly for the AVEC 2014 dataset. However, TDGAN heavily depends on visibly clear facial images.

A.4. Time and Memory Consumption

Once the optical lens is optimized and deployed, the inference time and memory consumption depend solely on the deep-learning model in OpticalDR. We compared FLOPs,



(a) ViT + lens

(b) STA-DRN (SA) + lens

(c) OpticalDR

Figure 1. CAM results for optical privacy-preserving DR approaches. Our (c) OpticalDR generates features with more variation compared to (a) ViT + lens, and exhibits clearer patterns compared to (b) STA-DRN + lens. Red areas represent high activation levels, while blue areas represent low activation levels.

Table 1. Performance Comparison of OpticalDR and TDGAN.

Method	Privacy-Preserving	AVEC 2013		AVEC 2014	
		MAE	RMSE	MAE	RMSE
TDGAN	×	8.56	10.75	8.26	10.50
OpticalDR	√	7.53	8.48	7.89	8.82

Table 2. Comparison of the time and memory consumption of the privacy-preserving DR model.

Method	FLOPs/G	Params/M	Time/ms
LQGDNet [5]	0.23	2.44	78.63
ViT [1]	11.29	58.07	4.15
STA-DRN [4]	4.84	33.32	2.63
OpticalDR	4.36	67.28	11.64

Params, and inference time in Tab. 2, and our method is comparable. It's crucial to note that the handcrafted feature extraction in LQGDNet makes its inference time not directly comparable with other deep-learning methods.

B. CAM Visualization

The full visualizing CAM results for OpticalDR and comparison with *ViT + lens* and *STA-DRN (SA) + lens* are shown in Fig. 1.

C. Visualization of Privacy Preserving Images

In this section, we present images generated by the lens within OpticalDR, showcasing samples with diverse levels of depression, as illustrated in Fig. 2. While the privacy-preserving images obscure identity details, discernible patterns persist. These patterns are likely intrinsic information utilized by the deep learning model for DR.

D. Visualization of Lens

We visualize the PSFs of the optimized lens in OpticalDR, as shown in Fig. 3.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv e-prints*, page arXiv:2010.11929, Oct. 2020. 2

[2] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 1

[3] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops (CVPRW)*, pages 94–101, 2010. 1

[4] Yuchen Pan, Yuanyuan Shang, Tie Liu, Zhuhong Shao, Guodong Guo, Hui Ding, and Qiang Hu. Spatial–temporal attention network for depression recognition from facial videos. *Expert Systems with Applications*, 237:121410, 2024. 2

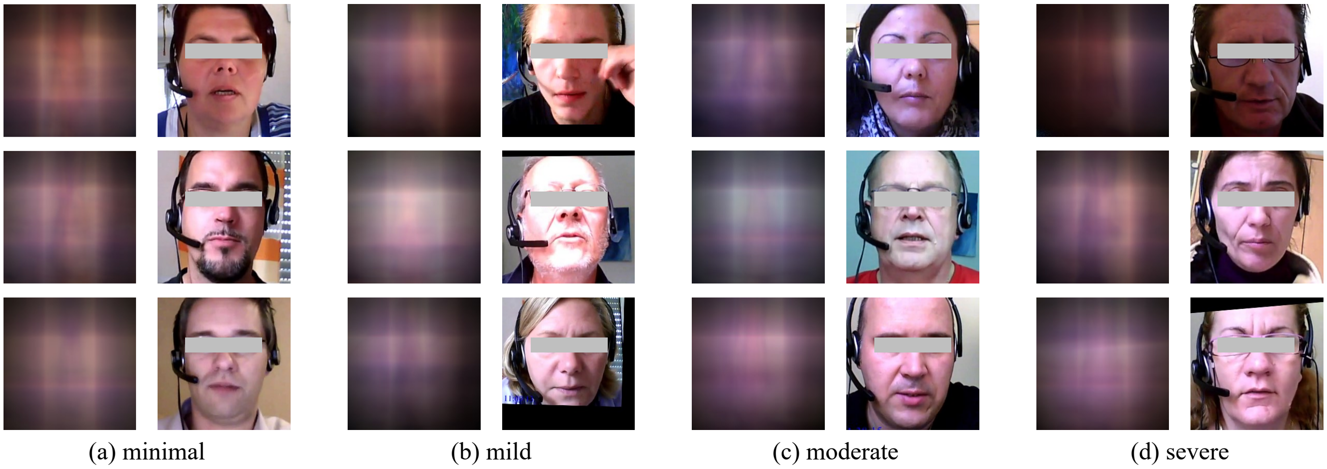

(a) minimal (b) mild (c) moderate (d) severe

Figure 2. The visualization of privacy-preserving images generated by OpticalDR at different depression levels. A high level of privacy preservation is achieved, making it difficult to observe individual information.

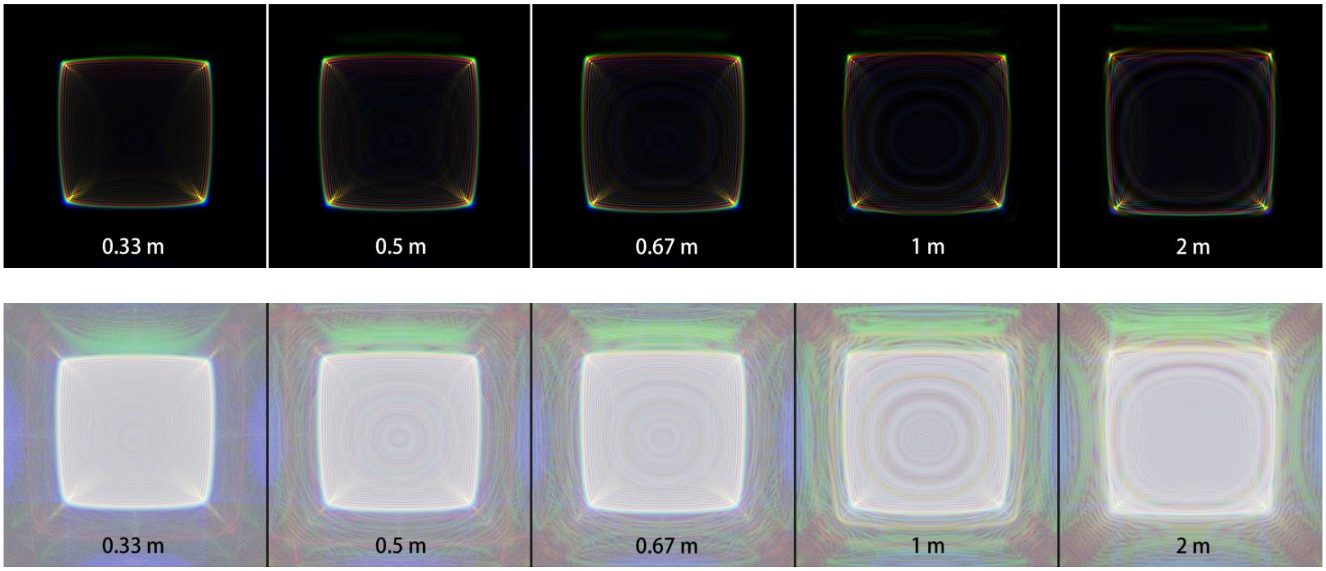


Figure 3. Visualization of simulated PSFs of the optimized lens in OpticalDR under various focal distances is shown in the top line, with log-transformed PSFs presented for enhanced clarity in the corresponding bottom line.

[5] Yuanyuan Shang, Yuchen Pan, Xiao Jiang, Zhhong Shao, Guodong Guo, Tie Liu, and Hui Ding. LQGDNet: A local quaternion and global deep network for facial depression recognition. *IEEE Transactions on Affective Computing*, pages 1–1, 2021. 2

[6] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6558–6569, 2019. 1

[7] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. AVEC 2014: 3D dimensional affect and depression recognition challenge. In *International Workshop on Audio/Visual Emotion Challenge (AVEC)*, page 3–10, 2014. 1

[8] Siyue Xie, Haifeng Hu, and Yizhen Chen. Facial expression recognition with two-branch disentangled generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2359–2371, 2021. 1