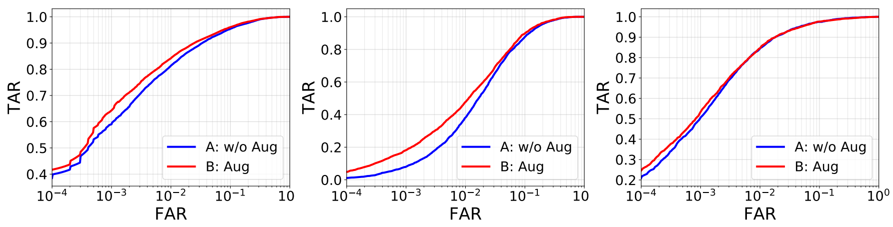
COSONet: Compact Second-Order Network for Video Face Recognition Supplemental Material

Yirong Mao[1,2][0000−0002−7603−4341], Ruiping Wang[1,2][0000−0003−1830−2595], Shiguang Shan[1,2][0000−0002−8348−392X], and Xilin Chen[1,2][0000−0003−3024−4404]

[1] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
{yirong.mao}@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn

1 Training Details

For the balancing scalar λ in the loss function, we tune it as in Table S1 and set it to be 0.01 in our all experiments. The scalar parameter $\mathbf{R}$ is initialized from the average L_2-norm of one batch of face descriptors in the first iteration, and then is updated by stochastic gradient descent (SGD). For training the ResNet-type network, the batch size is 256. The initial learning rate is 0.2 and decreased 0.5 per 3 epochs (training 36 epochs in total). For training the COSONet, the batch size is 128. The convolutional layers are initialized from its corresponding pre-trained ResNet-type network. The initial learning rate of the common layers is 0.001 and decreased 0.5 per 4 epochs. The initial learning rate of the newly layers is 0.05 and decreased 0.5 per 3 epochs. The total number of training epoch is 15. We train all networks using SGD with momentum of 0.9, and weight decay of $5e^{-4}$.



(a) IJB-A (b) PaSC-H (c) PaSC-C

Fig. S1. ROC curves for with or without data augmentation while training. 'w/o' represents the model without data augmentation.

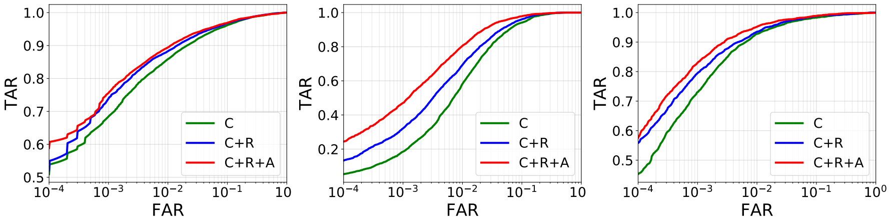
Table S1. The effect of λ. The corresponding model is Model II in Table 4. The reported performance is the TAR at FAR=$1e^{-3}$ on IJB-A dataset.

λ	0.001	0.005	0.01	0.05	0.1
TAR	0.724	0.731	0.735	0.723	0.711

2 More Experiment Results

The ROC curves for the effect of data augmentation are shown in Fig. S1. The network is ResNet-34 and the training dataset is WebFace. The ROC curves for the effect of ring loss are presented in Fig. S2. The network is COSONet based on ResNet-34 and the training dataset is WebFace.

It can be seen from Fig. S2 that data augmentation can upgrade the robustness of the face descriptors and improve the performance. Recently, a quite large video face dataset UMDFaces-Videos with 22,075 videos of 3,107 subjects (total 3.3M video frames) is published for training CNNs from scratch. This dataset has a certain number of overlapped subjects with VGGFace2 (about 541 subjects). In this paper, we train our COSONet on augmented VGGFace2 dataset, which has much facial variations, meanwhile, enjoys video-type noises. We have validated the effectiveness of data augmentation from the augmented VGGFace2 dataset. One can train CNNs on augmented VGGFace2 and UMDFaces-Videos simultaneously to further increase the performance. More detailed conclusions about the experiment results are discussed in our paper.



(a) IJB-A (b) PaSC-H (c) PaSC-C

Fig. S2. ROC curves for the impact of ring loss. 'C' represents the model with COSONet. 'A' represents the model trained with augmented faces. 'R' represents training with ring loss. Each technique is added gradually.