Supplementary Material: A Personalized Benchmark for Face Anti-spoofing

Davide Belli Debasmit Das Bence Major Fatih Porikli
Qualcomm AI Research*
{dbelli, debadas, bence, fporikli}@qti.qualcomm.com

A. Additional dataset details

A.1. SiW-Enroll folds

We create two different training and evaluation folds to make the face anti-spoofing task more challenging for SiW-Enroll. Fold 0 contains all live samples, high-resolution print attacks and iPad, iPhone replay attacks, while fold 1 includes live samples, low-resolution print attacks and replay attacks using Asus and Samsung devices. A model evaluated on test fold 0 is optimized on training fold 1, while we evaluate on test fold 1, models trained on training fold 0. In this way, the neural network will be trained on a subset of the available spoof mediums and evaluated on different ones, which is more realistic than assuming spoof mediums (like printed image and device resolution) to be unchanged at test time. When evaluating a method on SiW-Enroll, we report the average results over the two folds. In both SiW-Enroll and CelebA-Spoof-Enroll the set of test subject is disjoint from the set of training subjects.

A.2. Changing enrollment set size N

As discussed in Section 3 of the main paper, the proposed approach allows for the definition of personalized datasets with any number of enrollment images per query image. Indeed, In Section 5.2.3 we discuss experimental results obtained with enrollment sets of size $N = 8$. Depending on the original datasets, the population of the personalized version might change for different N values.

Table 5. Data statistics for the personalized benchmarks CASp-Enroll8, SiW-Enroll8.

	CASp-Enroll8		SiW-Enroll8	
	Train	Test	Train	Test
# data points	386,270	53,858	13,314	107,835
# subjects	7,021	851	90	75

For CelebA-Spoof, additional live images per subject are required to define the enrollment sets which reduces the number of query images for training and evaluation. Moreover, we discard subjects with less than N live images during the conversion process. This is done for simplicity but can be addressed by allowing enrollment sets of varying size $\leq N$. Notice how all aggregation methods except for concatenation support enrollment sets of different sizes.

In SiW, enrollment sets are generated by sampling equidistant frames from specific videos from the same subject. Since those videos are skipped when extracting frames for query images, changing the size of enrollment sets N does not impact the population of SiW-Enroll, as long as at least N frames are available in the reference video.

In Table 5 we report the dataset population for CelebA-Spoof-Enroll8 and SiW-Enroll8, showing that the available training and test data for the former dataset is reduced compared to CelebA-Spoof-Enroll5.

B. Implementation details

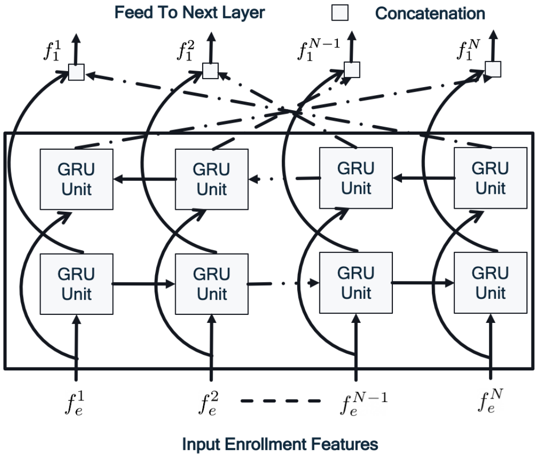

Figure 6. A single layer of the bidirectional GRU network learns to aggregate enrollment features by considering sequential relations.

To train a model on CelebA-Spoof-Enroll we use Adam [2] optimizer with a learning rate of 0.001 and apply weight decay with coefficient 0.001. The model is trained for 20k iterations using a batch size of 32. Following the original paper, the input images are resized to a size of 224×224 and color jitter augmentation is used with bright-

[*] Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

Table 6. Comparison between baseline and personalized model for the backbones VGG16, ResNet18 and FeatherNet on CASp-Enroll5 and SiW-Enroll5. The P-values are computed under the null hypothesis of "the personalized solution is producing worse or equal measurements compared to the baseline".

	CASp-Enroll5									**SiW-Enroll5**								
	VGG16			**ResNet18**			**FeatherNet**			**VGG16**			**ResNet18**			**FeatherNet**		
Method	**AUC**	**AUC10**	**EER**	**AUC**	**AUC10**	**EER**	**AUC**	**AUC10**	**EER**	**AUC**	**AUC10**	**EER**	**AUC**	**AUC10**	**EER**	**AUC**	**AUC10**	**EER**
Baseline	98.0	92.4	7.2	98.3	94.1	5.9	97.1	89.3	8.8	97.8	92.0	6.8	99.1	96.7	4.3	98.9	95.8	4.8
Personalized	**98.6**	**94.1**	**5.9**	**99.2**	**96.4**	**4.3**	**97.8**	**91.5**	**7.5**	**98.1**	**93.3**	**6.2**	**99.2**	**97.0**	**3.9**	**99.0**	**96.2**	**4.6**
P-value	.041	.050	.033	.015	.023	.019	$< .001$	$< .001$	.001	.123	.094	.105	.332	.198	.350	.211	.170	.150

ness, contrast, saturation and hue values of respectively 0, 0, 1 and 0.

For models trained on SiW-Enroll we use a similar configuration, with the difference of training for 10k iterations using a batch size of 128 and including an exponential decay every 1,000 iterations with $\gamma = 0.8$. The input images are resized to 128×128 pixels and no augmentation techniques are applied.

We follow the architecture description in the original papers to implement the backbones VGG16 [5], ResNet18 [1] and FeatherNet [6]. For the GRU implementation, we use $L = 2$ layers of bidirectional GRU to capture relationships in both temporal directions. The output of the bidirectional GRU layer at each step is thus a concatenation of the activations obtained from both the forward and reverse GRU at that step: $f_l^i = [h_l^i, \hat{h}_l^{N-i}]$. In Fig. 6 we show as an example the first GRU layer taking as input the sequence of enrollment images. For the attention-based method, the linear layers produce key and query features of size 64, while the values are 128-dimensional. We do not use dropout in the attention module. In the GNN-based method, we use three layers where each layer consists of the adjacency computation network and the graph convolution step. The architecture and hyper-parameter choice for the adjacency computation network are the same as the one used in [3].

The results for all sets of experiments are obtained with at least 5 different random seeds. For every seed, each metric is computed and averaged using 5 different model checkpoints from 0, 500, 1000, 1500 and 2000 steps before the end of the training. This helps ruling out experimental noise from the results and simplifying the comparison of different models.

C. Additional Experimental results

C.1. Statistical testing of experimental results

As means and standard deviations of results (without an agreed-upon comparison logic) only allow for informal estimates of significance, we employ statistical significance testing to gauge the strength of our conclusions. Also, the computation of standard deviation over different folds can lead to ambiguity in cases where the difference between fold difficulty is relatively high. In such cases, the standard deviation is biased to be higher than explained by the stochastic nature of the training process.

To measure the statistical significance of our main claim, we use the approximate randomization test for two independent samples, as described in [4]. The employed measure of difference between samples (from the sampling distribution) is the t-statistic from the t test for two independent samples, without assuming equal variances or samples sizes. Since computing the exact P-value from all possible arrangements is intractable, we use a sampling distribution of 10,000 random arrangements.

For P-values estimated over backbones or folds, we perform the re-sampling step in a way that measurements over different backbones or folds are not mixed and compared. A sample consists of the values used to measure the reported mean values (see Section B for details) for the given metric.

In Table 6 we provide the significance testing results for the main experiments discussed in the paper. Under the null hypothesis of "the personalized solution is producing worse or equal measurements compared to the baseline", we observe that P-values for all backbones applied to CASp-Enroll5 are lower than 0.05, confirming that personalization has a significant impact in improving anti-spoofing performance on this dataset. On SiW-Enroll5 we notice larger P-values, hinting towards personalization outperforming baselines in most cases, but not as consistently as for CASp-Enroll5.

Table 7. Results for baseline and personalized models aggregated over different backbones for SiW-Enroll5 and CASp-Enroll5.

	CASp-Enroll5			**SiW-Enroll5**		
# Enroll	**AUC**	**AUC10**	**EER**	**AUC**	**AUC10**	**EER**
Baseline	97.8	91.9	7.3	98.6	94.8	5.3
Personalized	**98.5**	**94.0**	**5.9**	**98.8**	**95.5**	**4.9**
P-value	.019	.024	.018	.222	.198	.157

In Table 7 we then report the results for personalized and baseline models averaged over the three backbones: VGG16, ResNet18 and FeatherNet. This is to summarize the expected impact of personalization for CelebA-Spoof-Enroll and SiW-Enroll datasets, abstracting away from the architectural choice. We also report significance testing results aggregated in the same way, which confirm the consistent effectiveness of personalization, especially for the CelebA-Spoof-Enroll dataset.

C.2. Number of enrollment images

In Table 8 we report the numerical results for the plot described in the main text evaluating the effect of changing the number of enrollment features. The takeaways are the same as the ones already described in the main text. While 2 and 4 enrollment images are optimal for the datasets we considered, we expect this hyper-parameter might be dataset-dependent, with more enrollment images being beneficial in case of larger variations across face images in the enrollment set.

Table 8. Effect of varying number of enrollment images out of a total of 8 when using VGG16 architecture on CASp-Enroll5 and SiW-Enroll5 datasets.

	CASp-Enroll5			**SiW-Enroll5**		
# Enroll	**AUC**	**AUC10**	**EER**	**AUC**	**AUC10**	**EER**
0/8	97.5	90.7	7.9	97.6	91.9	7.0
1/8	97.8	91.2	7.6	98.1	93.5	6.2
2/8	97.9	91.5	7.4	**98.2**	**94.2**	**6.0**
4/8	**98.1**	**92.6**	**6.8**	**98.2**	94.0	6.1
8/8	97.6	90.8	7.8	98.1	93.6	6.3

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[3] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *International Conference on Learning Representations*, 2018.

[4] David J Sheskin. *Handbook of parametric and nonparametric statistical procedures*. Chapman and Hall/CRC, 2003.

[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[6] Peng Zhang, Fuhao Zou, Zhiwen Wu, Nengli Dai, Skarpness Mark, Michael Fu, Juan Zhao, and Kai Li. Feathernets: convolutional neural networks as light as feather for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.