0.1 Models' specifications

We will compare the following four models against the TrustID Resnet-34 based solution, which comprises 29 layers, an average inference time of 7.28 seconds and has been trained on the VGG Face and Facescrub ensemble dataset using triplet loss. Relevant specifications that will assist in identifying the superior model are presented in Table 1. With Pytorch's model summary tool, the exact number of parameters, mult-adds and layers can be extracted, and from the models' documentation, the embedding size, loss used for training and the respective dataset.

	# Parameters	# Mult-Adds (G)	# Layers	Embedding	Inference time (s)	Loss	Dataset
MobileFaceNet	1,200,512	56.62	17	512	2.79	ArcFace	MS1MV2
iResnet-18	24,025,600	668.15	18	512	5.01	ArcFace	MS1MV2
FaceNet	28,907,599	152.21	63	512	5.89	Softmax	CASIA-WebFace
iResnet-SE-50	43,797,696	1610.00	50	512	9.78	ArcFace	MS1MV2

Table 1: Model's characteristics. "# Parameters" refers to the trainable parameters, "# Mult-Adds" to the number of multiplication and addition operations, "# Layers" denotes the quantity of convolutional and linear layers present in the model, "Embedding" signifies the dimensionality of the feature embedding produced by the model's output, "Inference time (s)" represents the average time taken for inference, "Loss" is the loss function used to train the model, and "Dataset" are the images employed for that action.

This initial analysis suggests that MobileFaceNet has promising characteristics. It is the least complex model, therefore, is less prone to overfitting to new data and, most importantly, the amount of computational overhead created and inferece times are much inferior to the others at study. This is supported by the fewer number of convolutional and linear layers, trainable parameters and mult-adds. Albeit the similar depth to that of iResnet-18, MobileFaceNet has, approximately, 95% fewer parameters, which is reflected on the number of total mult-adds. However, further investigation is required to determine whether the aforementioned characteristics might pose a bottleneck, potentially leading to a less robust solution with subpar performance. The ideal solution will strike a balance between adapting to new data and necessary computational costs.

0.2 Benchmarking Results

0.2.1 Accuracy

After performing 10-fold cross validation on all the benchmark datasets with the pretrained models, the mean accuracy is presented on the following table.

	Frontal		Age		Pose		Hard	
Models	CFP-FF	LFW	AgeDB30	CALFW	CFP-FP	CPLFW	VGGFace2	XQLFW
MobileFaceNet	0.9884	0.9912	0.9308	0.9362	0.8957	0.8642	0.9050	0.5063
iResnet-18	**0.9960**	**0.9960**	**0.9728**	**0.9555**	0.9414	0.8943	0.9198	0.4943
FaceNet	0.8909	0.9038	0.7147	0.7470	0.7664	0.6738	0.7748	0.5000
TrustID	0.8807	0.9906	0.7153	0.7198	0.7030	0.6235	0.7400	**0.6135**
iResnet-SE-50	0.9959	0.9953	0.9263	0.9543	**0.9457**	**0.9047**	**0.9396**	0.5137

Table 2: Model's face verification accuracy.

From Table 2, iResnet-18 achieves higher accuracy values on more datasets than any other model. However, iResnet-SE-50 performs better on the datasets where iResnet-18 falls short. Additionally, concerning the datasets where iResnet-SE-50 exhibited slightly lower performance, the accuracy scores are very close. Specifically, CFP-FF, CFP-FP, CALFW, and LFW are within a margin of error that can be attributed to non-deterministic behaviors in PyTorch, the libraries used, hardware, and/or CUDA. It is also important to note that, even though MobileFaceNet didn't achieve the higher accuracy on any benchmark, the scores are the third best and considering its lightweight specifications highlighted in Table 1, the results are very promising and present a good example of accuracy and computational cost trade-off. Regarding the XQLFW results, with the exception of TrustID, they are exceedingly low, approaching 0.5 or worse. This suggests that the model is producing outputs that resemble random guesses, which is exactly what occurs with FaceNet.

0.2.2 ROC Curves

The previous table indicates that the three best performers, based exclusively on the accuracy at the best similarity threshold for each model and dataset, are the iResnet-SE-50, iResnet-18 and MobileFaceNet. To conduct a more thorough investigation allowing us to select the best model to be fine-tuned, the TAR values are calculated for a range of FAR values and the ROC curves are generated. By inspecting how close the ROC curve is to the top left corner, the best models can be determined since those are able to correctly identify more genuine matches while keeping the impostor matches low.

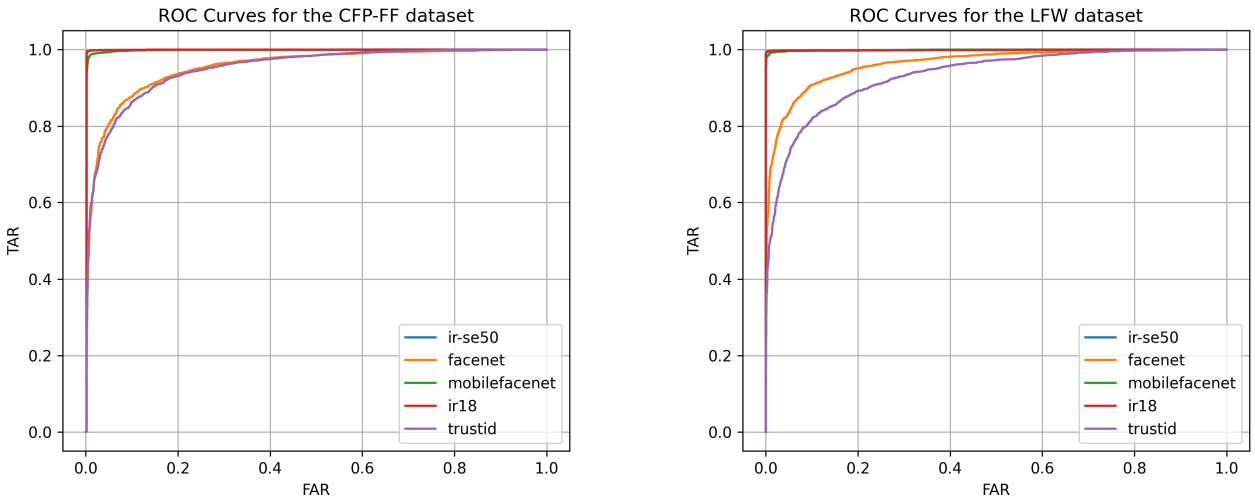

Figure 1: ROC Curves for the frontal pose face verification datasets.

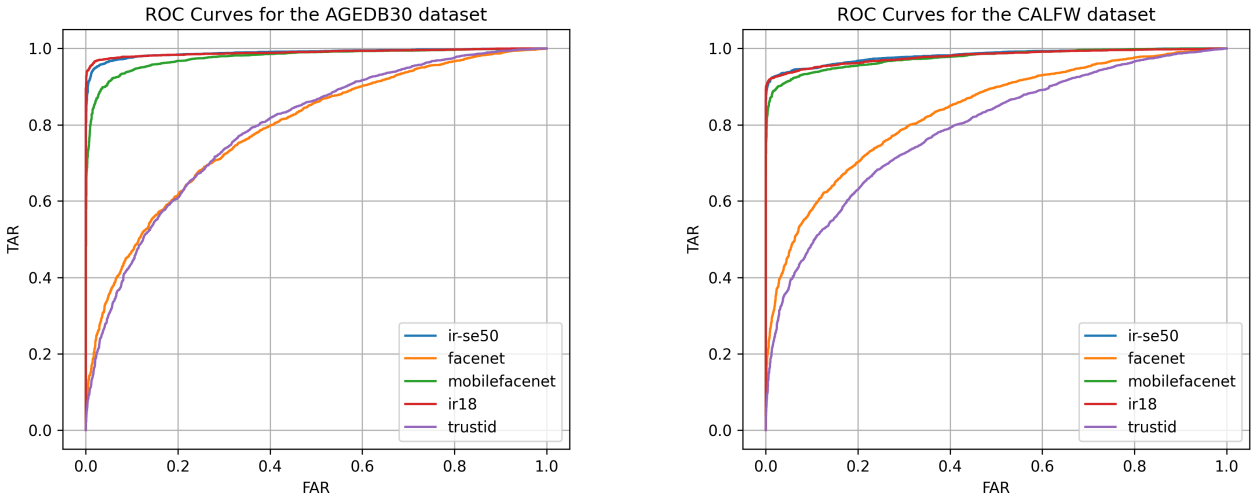

Figure 2: ROC Curves for the age face verification datasets.

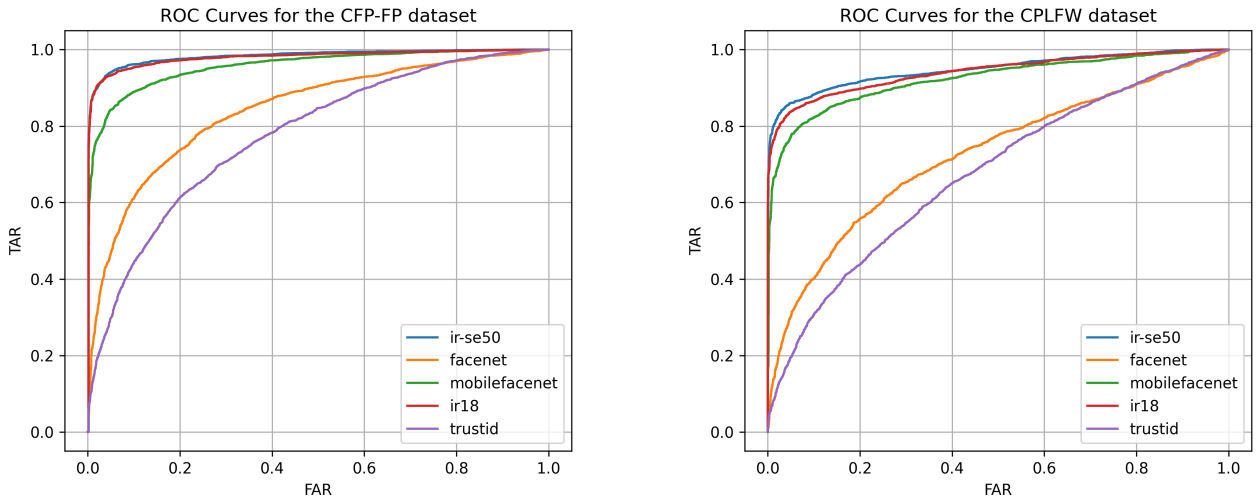

Figure 3: ROC Curves for the pose face verification datasets.

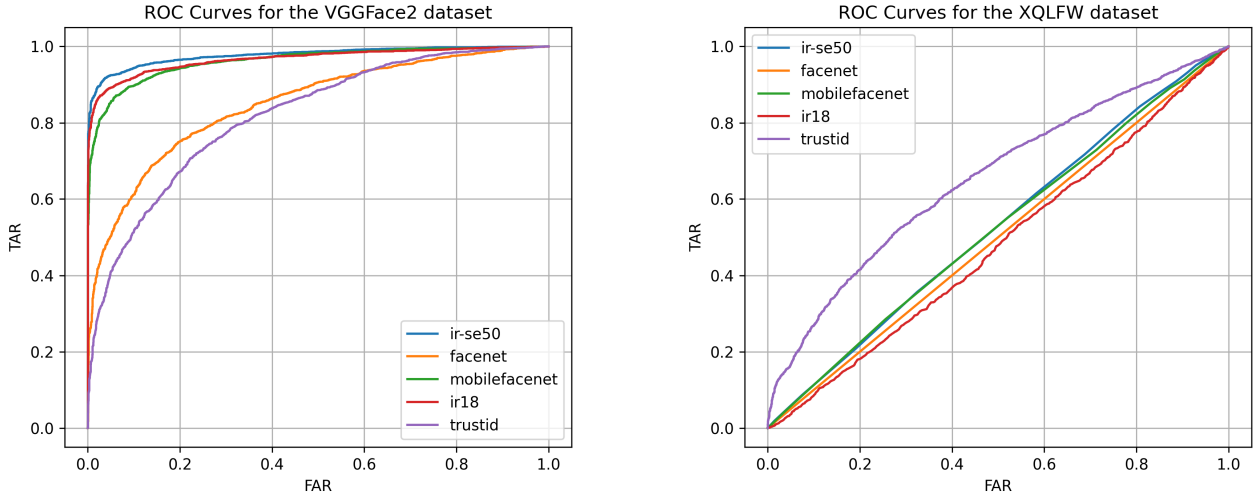

Figure 4: ROC Curves for the hard face verification datasets.

The ROC curves support our initial assumptions. It is evident that, across the entire FAR range, the three top-performing models are iResnet-SE-50, iResnet-18, and MobileFacenet. With the exception of XQLFW in (Figure 4), where all models, except TrustID, perform poorly and are close to random guessing, the remaining seven datasets consistently position these three models near the top-left corner. This pattern indicates strong model performance. In scenarios with low FAR values, where the model is less tolerant of incorrectly identifying impostors as matches, the number of correctly classified pairs (TAR) is higher.

		iResnet-SE-50			iResnet-18		MobileFacenet			FaceNet		TrustID	
		$1e-4$	$1e-3$	1e-2	$1e-3$	$1e-2$	$1e-4$	$1e-3$	$1e-2$	$1e-3$	$1e-2$	$1e-3$	$1e-2$
Frontal	CFP-FF	0.10000	0.10000	0.99771	0.09886	0.99743	0.09057	0.09600	0.98657	0.02886	0.55257	0.03314	0.58114
	LFW	0.99067	0.99333	0.99700	0.99100	0.99600	0.91467	0.96933	0.99167	0.39900	0.68800	0.26967	0.50567
Age	AgeDB30	0.68267	0.81700	0.92600	0.91533	0.95300	0.4950	0.59500	0.80267	0.02900	0.14867	0.03467	0.11100
	CALFW	0.86633	0.88233	0.91733	0.90167	0.91933	0.68100	0.75900	0.87100	0.07867	0.26767	0.05300	0.18267
Pose	CFP-FP	0.07600	0.07971	0.87371	0.07628	0.87400	0.04200	0.05171	0.69857	0.00857	0.22171	0.00686	0.12629
	CPLFW	0.37533	0.58833	0.7900	0.66767	0.75400	0.06467	0.17200	0.64400	0.00967	0.12433	0.01567	0.07300
Hard	VGGFace2	0.06000	0.77280	0.86280	0.70320	0.81840	0.05240	0.53880	0.72160	0.17000	0.31720	0.06640	0.19400
	XQLFW	0.00001	0.00033	0.00800	0.00033	0.00433	0.00001	0.00100	0.00433	0.02000	0.40433	0.02167	0.07867

Table 3: TAR@FAR for all the models and benchmarks

The aforementioned three highest-achieving models distance themselves from FaceNet and TrustID on the ROC plots, although there is some overlap. Therefore, Table 3 allows us to analyze their performance at lower FAR values, where this overlap occurs.

For $FAR = 1e-4$, all models fail on more demanding datasets, but for easier ones iResnet-SE-50 is capable of good performance on LFW and CALFW, and MobileFaceNet on LFW. Reducing the strictness and increasing the FAR to $1e-3$, there's an improvement on the results as expected. The iResnet models produce high TAR values on all datasets apart from the more challenging CFP variations and XQLFW datasets, MobileFacenet starts to improve but still performs poorly on the pose group, hard group, CFP pair and AgeDB30 dataset. Finally, at $1e-2$ is the threshold at which all models excel without compromising the security of the system, since increasing the FAR to $1e-1$ would lead to too much falsely matched pairs. iResnet-SE-50 and iResnet-18 have comparable performance with high scores on the same benchmarks and both failing XQLFW. MobileFaceNet approaches iResnet levels of capability aside from slightly lower scores on the age and pose groups and the usual XQLFW dataset.

0.2.3 DET Curves

To finalize the choice of the best model, the FRRs are calculated and plotted against the previous FAR values to obtain the DET curves. The intersection between the identity line that divides the graph and the DETs, the EER points can be extracted. These curves also allow to make a better distinction between models due to the more expansive logarithmic scale in which they are generated. In this case, contrary to the ROC curves, the better performing models are closer to the lower left corner, since that will minimize both the amount of impostors matched as true identities (FAR) and true identities classified as impostors (FRR).

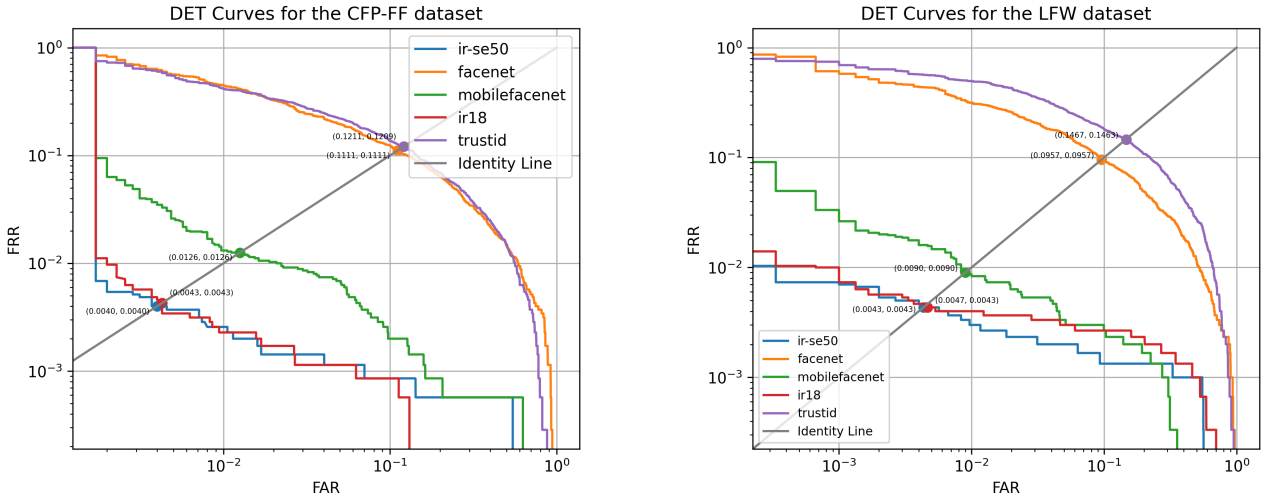

Figure 5: DTE Curves and EER points for the frontal pose face verification datasets.

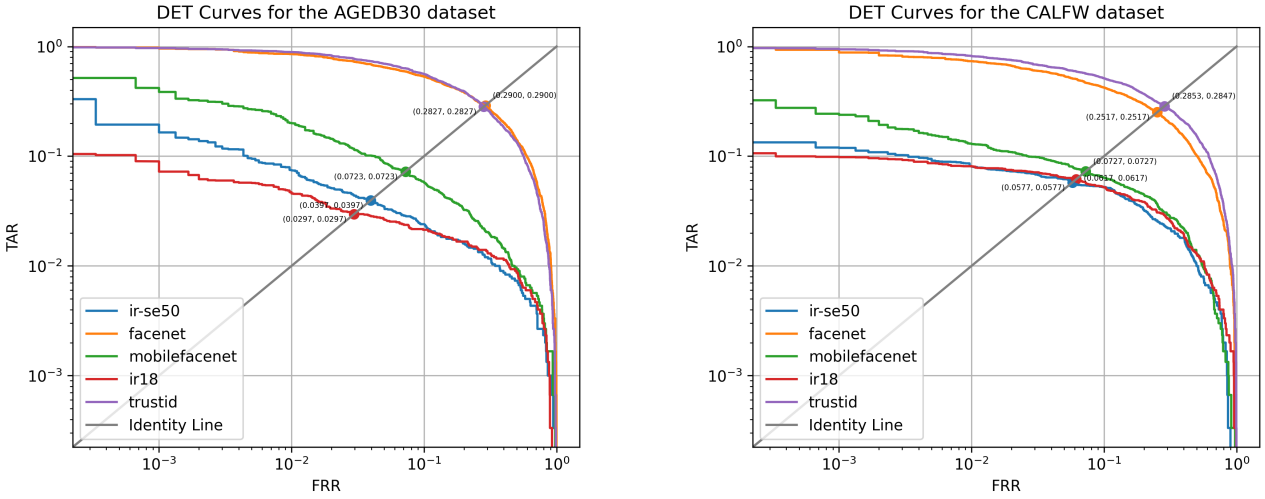

Figure 6: DET Curves and EER points for the age face verification datasets.

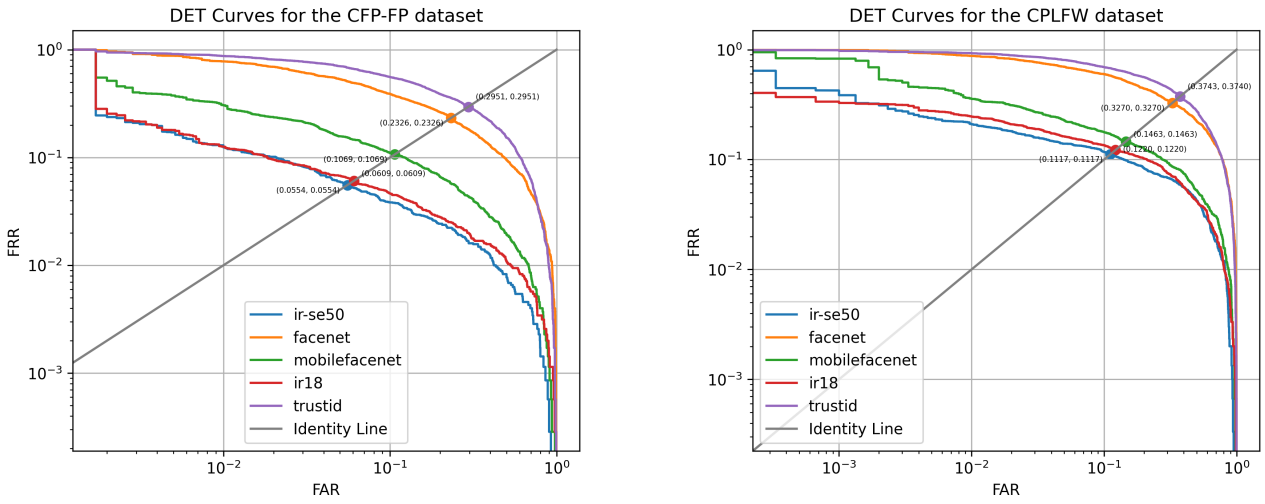

Figure 7: DET Curves and EER points for the pose face verification datasets.

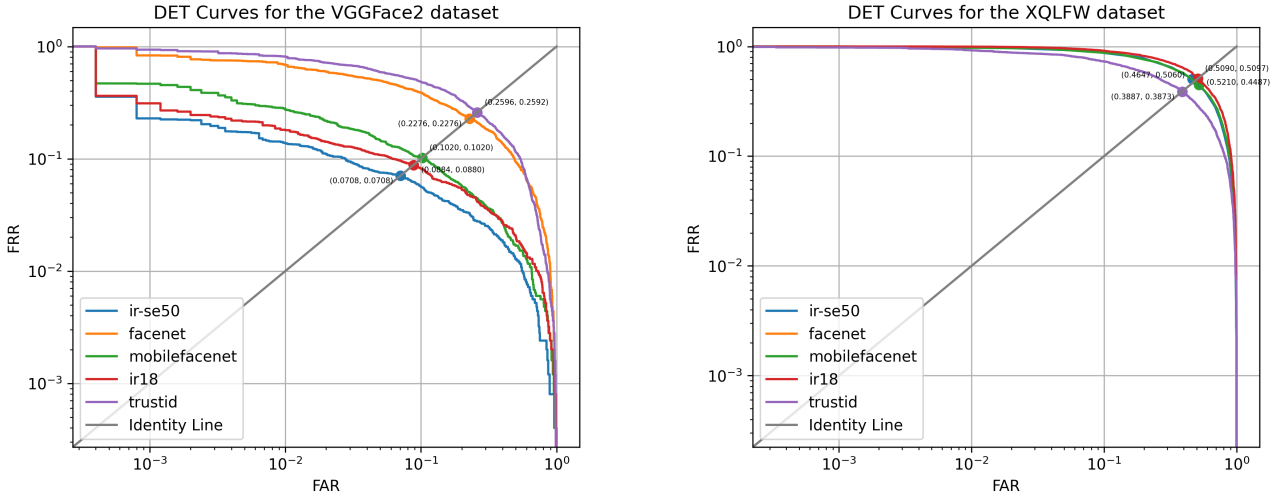

Figure 8: DET Curves and EER points for the hard face verification datasets.

	Frontal		Age		Pose		Hard	
	CFP-FF	LFW	AgeDB30	CALFW	CFP-FP	CPLFW	VGGFace2	XQLFW
MobileFaceNet	0.0126	0.0090	0.0723	0.0727	0.1069	0.1463	0.1020	0.4849
iResnet-18	0.0043	0.0045	**0.0297**	0.0617	0.0609	0.1220	0.0882	0.5094
FaceNet	0.1111	0.0957	0.2900	0.2517	0.2326	0.3270	0.2276	-
TrustID	0.1210	0.1465	0.2827	0.2850	0.2951	0.3742	0.2594	**0.3880**
iResnet-SE-50	**0.0040**	**0.0043**	0.0723	**0.0577**	**0.0554**	**0.1117**	**0.0708**	0.4854

Table 4: EER values for all the models and respective benchmarks.

As a final analysis, both the DET plots and the EER values support what has been previously discussed: iResnet-SE-50, iResnet-18 and MobileFaceNet are the best performing models. To no surprise, this group is always close to each other and close to the ideal corner of the DET graphs. Additionally, XQLFW reveals once again to be too much of a challenge. Regarding the EER scores, the results and conclusions are similar to the ones from Table 2. iResnet-SE-50, the bigger and more complex value, has the lower values, with iResnet-18 a close second and MobileFaceNet the third best.

0.2.4 Discussion

At this stage, we are capable of assuring, based on the previous tests, that there are better answers to TrustID's facial verification framework. iResnet-SE-50 is the overall best achieving model, iResnet-18 is the second best and MobileFaceNet is just slightly worse than the very close aforementioned two. The top three highlighted throughout prior sections proved their robustness to extreme variation in pose (CFP-FP and CPLFW), age (AgeDB30 and CALFW) or illumination (CFP-FF, VGGFace2). That being said, the benchmark concerning quality and image degradation (XQLFW) proved to be a major hurdle to most of the models aside from TrustID. That can be justified by the method of training. By resizing smaller training images to 150×150 there's a degradation in quality that leads to a model more prepared to handle these situations.

Considering the benchmark's results, the much lower amount of trainable parameters, mult-adds and inference time, MobileFaceNet is the best balance between performance and computational cost. Furthermore, although it is true that Mobilefacenet's accuracy is high in the pose group and VGGFace2, there's room for improvement in the TAR at low FAR values, hence the fine-tuning. The objective of further training the model first with QMUL-SurvFace is to try and improve its scores in both the hard group, but specially XQLFW, and pose group benchmarks. On the other hand, DigiFace-1M enables the study of the impact of fully-synthetic ethically collected data on the scores throughout the benchmarks, with a special attention to the pose group benchmarks.

0.3 Training Details

By leveraging Optuna's hyperparameter searching capabilities, we concluded that the optimal combination is to train over batches of size 32 for 10 epochs with a $1e-4$ learning rate that decays according to a Cosine Annealing scheduler with warm up restarts. Additionally, because the model has Batch Normalization layers, they need to be explicitly set to evaluation mode during training. If this

step is overlooked, the mean and variance used will be the ones from the batch and not the values achieved during the pre-training, leading to bad evaluation values. Following the original ArcFace paper [1], the optimizer of choice is Stochastic Gradient Descent with momentum 0.9 and weight decay $5e-4$, scale $s = 64$ and margin $m = 0.5$. Early stopping, by evaluating on XQLFW and CPLFW during training, is also adopted as a safety measure to guarantee the best results possible.

0.4 Training Results

The following sections will gather the results from the several experiments that were performed in order to evaluate what solution is a better fit for the initial problem.

0.4.1 Whole network tuning

The first approach is to update the whole network, because both QMUL-SurvFace and DigiFace-1M are big datasets with a great number of images and identities, therefore, there are less chances of quickly overfitting to the training data.

Benchmarks		Original	$m = 0.5$	$m = 0.4$	$m = 0.3$
Frontal	CFP-FF	0.9884	0.7957 (↓)	0.7916 (↓)	0.7916 (↓)
	LFW	0.9912	0.7957 (↓)	0.7915 (↓)	0.7913 (↓)
Age	AgeDB30	0.9308	0.6165 (↓)	0.6112 (↓)	0.6362 (↓)
	CALFW	0.9362	0.6593 (↓)	0.6235 (↓)	0.6472 (↓)
Pose	CFP-FP	0.8957	0.6447 (↓)	0.6381 (↓)	0.6556 (↓)
	CPLFW	0.8642	0.6048 (↓)	0.5843 (↓)	0.5992 (↓)
Hard	VGGFace2	0.9050	0.6516 (↓)	0.6188 (↓)	0.6314 (↓)
	XQLFW	0.5063	0.5325 (↑)	0.5355 (↑)	0.5215 (↑)
Stopping Epoch			6	7	7

Table 5: MobileFaceNet accuracies before and after being fine-tuning the whole network on QMUL-SurvFace with different ArcFace margins.

According to Table 5, fine-tuning with QMUL-SurvFace improves, as would be expected, the XQLFW benchmark performance by 5.17%. However, the verified increase is moderate and disadvantageous when the accuracy degradation verified on the remaining benchmarks is taken into account.

In an effort to improve the results, the margin is reduced in order to generate a less penalizing training with a more lax distance between classes. The results

from this settings are also shown in Table 5, and when fine-tuned with QMUL-SurvFace, the XQLFW results are better than the original pre-trained model, increasing 5.77% for $m = 0.4$ and 3.00% for $m = 0.3$.

Benchmarks		Original	$m = 0.5$	$m = 0.4$	$m = 0.3$
Frontal	CFP-FF	0.9884	0.8011 (↓)	0.8231 (↓)	0.8157 (↓)
	LFW	0.9912	0.8011 (↓)	0.8231 (↓)	0.8157 (↓)
Age	AgeDB30	0.9308	0.6732 (↓)	0.6733 (↓)	0.6723 (↓)
	CALFW	0.9362	0.6755 (↓)	0.7010 (↓)	0.7010 (↓)
Pose	CFP-FP	0.8957	0.6191 (↓)	0.6483 (↓)	0.6609 (↓)
	CPLFW	0.8642	0.5945 (↓)	0.6078 (↓)	0.6170 (↓)
Hard	VGGFace2	0.9050	0.6520 (↓)	0.6744 (↓)	0.6758 (↓)
	XQLFW	0.5063	0.4965 (↓)	0.5003 (↓)	0.4975 (↓)
Stopping Epoch			6	5	6

Table 6: MobileFaceNet accuracies before and after being fine-tuning the whole network on DigiFace-1M with different ArcFace margins.

In the case of DigiFace-1M Table 6, no discernible improvements were observed, and reducing the margin size did not yield any positive changes. Instead, the model appeared to struggle in adapting to the dataset, resulting in adjustments to the weights that ultimately led to a deterioration in benchmark performance.

When comparing directly with the paper's suggested margin ($m = 0.5$), some conclusions can be drawn. The model fine-tuned with QMUL-SurvFace produces worse results in the frontal and age groups for $m = 0.4$ and $m = 0.3$, the pose group is better for both margins and the hard group is mixed, where VGGFace2 has worse performance for $m = 0.4$ and $m = 0.3$, and XQLFW improves for $m = 0.4$ and not for $m = 0.3$. On the other hand, when $m = 0.4$ and $m = 0.3$, the DigiFace-1M training saw a marginal improvement throughout the tests but still performs poorly. All in all, reducing the margin size does not have a meaningful impact on the accuracy results.

To achieve a more profound understanding of how the models reacts to the data, Table 7 and Table 8 present the TAR at very low FAR.

		$m = 0.5$			$m = 0.4$			$m = 0.3$		
Benchmarks		$1e-4$	$1e-3$	$1e-2$	$1e-4$	$1e-3$	$1e-2$	$1e-4$	$1e-3$	$1e-2$
Frontal	CFP-FF	0.0140 (↓)	0.0209 (↓)	0.3486 (↓)	0.0063 (↓)	0.0154 (↓)	0.2940 (↓)	0.0091 (↓)	0.0169 (↓)	0.3097 (↓)
	LFW	0.3099 (↓)	0.3263 (↓)	0.4760 (↓)	0.1267 (↓)	0.1680 (↓)	0.3940 (↓)	0.1927 (↓)	0.2310 (↓)	0.4697 (↓)
Age	AgeDB30	0.0050 (↓)	0.0083 (↓)	0.0867 (↓)	0.0027 (↓)	0.0053 (↓)	0.0730 (↓)	0.0093 (↓)	0.0240 (↓)	0.0860 (↓)
	CALFW	0.0313 (↓)	0.8053 (↑)	0.8053 (↓)	0.0070 (↓)	0.0323 (↓)	0.0827 (↓)	0.0283 (↓)	0.0340 (↓)	0.0920 (↓)
Pose	CFP-FP	0.0006 (↓)	0.9380 (↑)	0.9380 (↑)	0.0003 (↓)	0.0003 (↓)	0.0666 (↓)	0.0000 (↓)	0.0003 (↓)	0.0626 (↓)
	CPLFW	0.0060 (↓)	0.9664 (↑)	0.9664 (↑)	0.0050 (↓)	0.0067 (↓)	0.0397 (↓)	0.0073 (↓)	0.0147 (↓)	0.0613 (↓)
Hard	VGGFace2	0.0040 (↓)	0.0536 (↓)	0.1060 (↓)	0.0036 (↓)	0.0376 (↓)	0.0880 (↓)	0.0040 (↓)	0.0364 (↓)	0.1216 (↓)
	XQLFW	0.0000 (↓)	0.0023 (↑)	0.0173 (↑)	0.0000 (↓)	0.0000 (↓)	0.0000 (↓)	0.0000 (↓)	0.0000 (↓)	0.0000 (↓)

Table 7: TAR@FAR after fine-tuning the model with QMUL-SurvFace.

Benchmarks		$m = 0.5$			$m = 0.4$			$m = 0.3$		
		$1e-4$	$1e-3$	$1e-2$	$1e-4$	$1e-3$	$1e-2$	$1e-4$	$1e-3$	$1e-2$
Frontal	CFP-FF	0.0020 (↓)	0.0069 (↓)	0.1906 (↓)	0.0026 (↓)	0.0111 (↓)	0.3580 (↓)	0.0057 (↓)	0.0109 (↓)	0.3606 (↓)
	LFW	0.1633 (↓)	0.2810 (↓)	0.5520 (↓)	0.2300 (↓)	0.3403 (↓)	0.6310 (↓)	0.2320 (↓)	0.3743 (↓)	0.6113 (↓)
Age	AgeDB30	0.0013 (↓)	0.0083 (↑)	0.0530 (↓)	0.0010 (↓)	0.0153 (↓)	0.0747 (↓)	0.0003 (↓)	0.0250 (↓)	0.0790 (↓)
	CALFW	0.0017 (↓)	0.0043 (↓)	0.0780 (↓)	0.0013 (↓)	0.0300 (↓)	0.1207 (↓)	0.0050 (↓)	0.0170 (↓)	0.1223 (↓)
Pose	CFP-FP	0.0000 (↓)	0.0000 (↓)	0.0149 (↓)	0.0000 (↓)	0.0000 (↓)	0.0194 (↓)	0.0000 (↓)	0.0003 (↓)	0.0226 (↓)
	CPLFW	0.0013 (↓)	0.0017 (↓)	0.0190 (↓)	0.0007 (↓)	0.0016 (↓)	0.0263 (↓)	0.0007 (↓)	0.0033 (↓)	0.0287 (↓)
Hard	VGGFace2	0.0000 (↓)	0.0008 (↓)	0.0220 (↓)	0.0000 (↓)	0.0020 (↓)	0.0404 (↓)	0.0000 (↓)	0.0028 (↓)	0.0440 (↓)
	XQLFW	0.0000 (↓)	0.0000 (↓)	0.0083 (↓)	0.0003 (↑)	0.0013 (↑)	0.0087 (↑)	0.0000 (↓)	0.0010 (–)	0.0073 (↑)

Table 8: TAR@FAR after fine-tuning the model with DigiFace-1M.

In the context of general CNN architecture, it's well-established that the initial layers are primarily responsible for learning fundamental features such as edges, basic shapes, and patterns that constitute objects or faces. Therefore, with the intention of improving the previous results, by preserving the weights associated with these earlier layers and avoiding introducing noise during further training, two other approaches are made: 1) freeze the first 5 layers and 2) train only the 2 final layers. Additionally, following the logic from the previous experience, different ArcFace margins are also studied.

0.4.2 5 layers

Benchmarks		Original	$m = 0.5$	$m = 0.4$	$m = 0.3$
Frontal	CFP-FF	0.9884	0.7821 (↓)	0.8414 (↓)	0.8155 (↓)
	LFW	0.9912	0.7820 (↓)	0.8415 (↓)	0.8156 (↓)
Age	AgeDB30	0.9308	0.5985 (↓)	0.6373 (↓)	0.6382 (↓)
	CALFW	0.9362	0.6506 (↓)	0.6962 (↓)	0.6765 (↓)
Pose	CFP-FP	0.8957	0.6477 (↓)	0.6627 (↓)	0.6499 (↓)
	CPLFW	0.8642	0.5943 (↓)	0.6195 (↓)	0.5872 (↓)
Hard	VGGFace2	0.9050	0.6442 (↓)	0.6822 (↓)	0.6610 (↓)
	XQLFW	0.5063	0.4925 (↓)	0.5020 (↓)	0.5127 (↑)
Stopping Epoch			8	5	6

Table 9: MobileFaceNet accuracies before and after being fine-tuning the network, with the first five layers frozen, on QMUL-SurvFace with different ArcFace margins.

Benchmarks		Original	$m = 0.5$	$m = 0.4$	$m = 0.3$
Frontal	CFP-FF	0.9884	0.8840 (↓)	0.8806 (↓)	0.8788 (↓)
	LFW	0.9912	0.8840 (↓)	0.8805 (↓)	0.8789 (↓)
Age	AgeDB30	0.9308	0.7265 (↓)	0.7125 (↓)	0.7303 (↓)
	CALFW	0.9362	0.7400 (↓)	0.7478 (↓)	0.7398 (↓)
Pose	CFP-FP	0.8957	0.7219 (↓)	0.7039 (↓)	0.7223 (↓)
	CPLFW	0.8642	0.6423 (↓)	0.6335 (↓)	0.6435 (↓)
Hard	VGGFace2	0.9050	0.7220 (↓)	0.7122 (↓)	0.7080 (↓)
	XQLFW	0.5063	0.4997 (↓)	0.4967 (↓)	0.5033 (↓)
Stopping Epoch			6	6	6

Table 10: MobileFaceNet accuracies before and after fine-tuning the network, with the first five layers frozen, on DigiFace-1M with different ArcFace margins.

		m=0.5			m=0.4			m=0.3		
Benchmarks		1e-4	1e-3	1e-2	1e-4	1e-3	1e-2	1e-4	1e-3	1e-2
Frontal	CFP-FF	0.0117 (↓)	0.0149 (↓)	0.2769 (↓)	0.0174 (↓)	0.0303 (↓)	0.4786 (↓)	0.0163 (↓)	0.0286 (↓)	0.3897 (↓)
	LFW	0.2033 (↓)	0.2277 (↓)	0.5047 (↓)	0.2907 (↓)	0.3473 (↓)	0.5870 (↓)	0.2597 (↓)	0.2743 (↓)	0.5533 (↓)
Age	AgeDB30	0.0137 (↓)	0.0207 (↓)	0.0597 (↓)	0.0087 (↓)	0.0180 (↓)	0.0643 (↓)	0.0063 (↓)	0.0177 (↓)	0.0777 (↓)
	CALFW	0.018 (↓)	0.0210 (↓)	0.0877 (↓)	0.0667 (↓)	0.0737 (↓)	0.1510 (↓)	0.0267 (↓)	0.0570 (↓)	0.1450 (↓)
Pose	CFP-FP	0.0000 (↓)	0.0009 (↓)	0.0617 (↓)	0.0011 (↓)	0.0029 (↓)	0.0871 (↓)	0.0009 (↓)	0.0020 (↓)	0.0694 (↓)
	CPLFW	0.0070 (↓)	0.0170 (↓)	0.0570 (↓)	0.0100 (↓)	0.0183 (↓)	0.0870 (↓)	0.0093 (↓)	0.0187 (↓)	0.0633 (↓)
Hard	VGGFace2	0.0032 (↓)	0.0352 (↓)	0.1112 (↓)	0.0056 (↓)	0.0712 (↓)	0.1516 (↓)	0.0040 (↓)	0.0528 (↓)	0.1304 (↓)
	XQLFW	0.0000 (↓)	0.0000 (↓)	0.0000 (↓)	0.0000 (↓)	0.0000 (↓)	0.0040 (↓)	0.0000 (↓)	0.0003 (↓)	0.0037(↓)

Table 11: TAR@FAR after fine-tuning the model, with the first five layers frozen, on QMUL-SurvFace.

		m=0.5			m=0.4			m=0.3		
Benchmarks		1e-4	1e-3	1e-2	1e-4	1e-3	1e-2	1e-4	1e-3	1e-2
Frontal	CFP-FF	0.0189 (↓)	0.0354 (↓)	0.5574 (↓)	0.0183 (↓)	0.0303 (↓)	0.5200 (↓)	0.0294 (↓)	0.0346 (↓)	0.5539 (↓)
	LFW	0.3207 (↓)	0.5577 (↓)	0.7440 (↓)	0.2677 (↓)	0.5180 (↓)	0.7360 (↓)	0.4420 (↓)	0.5980 (↓)	0.7497 (↓)
Age	AgeDB30	0.0083 (↓)	0.0363 (↓)	0.1450 (↓)	0.0027 (↓)	0.0203 (↓)	0.1407 (↓)	0.0047 (↓)	0.0253 (↓)	0.1353 (↓)
	CALFW	0.0407 (↓)	0.0813 (↓)	0.2690 (↓)	0.0100 (↓)	0.0593 (↓)	0.2423 (↓)	0.0150 (↓)	0.0393 (↓)	0.2590 (↓)
Pose	CFP-FP	0.0006 (↓)	0.0026 (↓)	0.1231 (↓)	0.0003 (↓)	0.0029 (↓)	0.0880 (↓)	0.0003 (↓)	0.0017 (↓)	0.1106 (↓)
	CPLFW	0.0003 (↓)	0.0017 (↓)	0.0383 (↓)	0.0007 (↓)	0.0023 (↓)	0.0247 (↓)	0.0003 (↓)	0.0017 (↓)	0.0407 (↓)
Hard	VGGFace2	0.0000 (↓)	0.0176 (↓)	0.1732 (↓)	0.0000 (↓)	0.0056 (↓)	0.1392 (↓)	0.0000 (↓)	0.0040 (↓)	0.1468 (↓)
	XQLFW	0.0000 (↓)	0.0010 (−)	0.0100 (↑)	0.0003 (↑)	0.0007 (↓)	0.0067 (↑)	0.0000 (↓)	0.0010 (−)	0.0093 (↑)

Table 12: TAR@FAR after fine-tuning the model, with the first five layers frozen, on DigiFace-1M.

0.4.3 Final Layers

Benchmarks		Original	$m = 0.5$	$m = 0.4$	$m = 0.3$
Frontal	CFP-FF	0.9884	0.8751 (↓)	0.8821 (↓)	0.8786 (↓)
	LFW	0.9912	0.8751 (↓)	0.8821 (↓)	0.8785 (↓)
Age	AgeDB30	0.9308	0.7097 (↓)	0.7220 (↓)	0.7080 (↓)
	CALFW	0.9362	0.7762 (↓)	0.7833 (↓)	0.7752 (↓)
Pose	CFP-FP	0.8957	0.6729 (↓)	0.6714 (↓)	0.6719 (↓)
	CPLFW	0.8642	0.6585 (↓)	0.6635 (↓)	0.6635 (↓)
Hard	VGGFace2	0.9050	0.6992 (↓)	0.7014 (↓)	0.7006 (↓)
	XQLFW	0.5063	0.4975 (↓)	0.4993 (↓)	0.5010 (↓)
Stopping Epoch			6	3	6

Table 13: MobileFaceNet accuracies before and after fine-tuning the network, with all the layers frozen aside the last two, on QMUL-SurvFace with different ArcFace margins.

Benchmarks		Original	$m = 0.5$	$m = 0.4$	$m = 0.3$
Frontal	CFP-FF	0.9884	0.9568 (↓)	0.9630 (↓)	0.9574 (↓)
	LFW	0.9912	0.9569 (↓)	0.9629 (↓)	0.9574 (↓)
Age	AgeDB30	0.9308	0.8420 (↓)	0.8533 (↓)	0.8403 (↓)
	CALFW	0.9362	0.8672 (↓)	0.8795 (↓)	0.8643 (↓)
Pose	CFP-FP	0.8957	0.8133 (↓)	0.8199 (↓)	0.8171 (↓)
	CPLFW	0.8642	0.7650 (↓)	0.7830 (↓)	0.7697 (↓)
Hard	VGGFace2	0.9050	0.8158 (↓)	0.8290 (↓)	0.8134 (↓)
	XQLFW	0.5063	0.4923 (↓)	0.4995 (↓)	0.4993 (↓)
Stopping Epoch			5	3	4

Table 14: MobileFaceNet accuracies before and after fine-tuning the network, with all the layers frozen aside the last two, on DigiFace-1M with different ArcFace margins.

?? and **??** present the results for the aforementioned experiences. For both QMUL-SurvFace and DigiFace-1M, there are no improvements at any margin value and the XQLFW improvements are lost with the exception of $m = 0.3$. The only observable difference is the reduced deterioration of the original weights, and although it produces better results than the first experience where the whole network is trained, they are still worse than the pretrained model. A common pattern between the three training approaches is that, for $m = 0.4$, the model appears to reach its peak accuracy, and for both the partial trainings, that value is reached faster, resulting in less weight deterioration. That behavior is foreseen, since the model trains for less epochs, which leads the weights to be updated less times, preventing them from getting worse.

		m=0.5			m=0.4			m=0.3		
Benchmarks		1e-4	1e-3	1e-2	1e-4	1e-3	1e-2	1e-4	1e-3	1e-2
Frontal	CFP-FF	0.038 (↓)	0.0394 (↓)	0.6503 (↓)	0.0280 (↓)	0.0369 (↓)	0.6237 (↓)	0.0306 (↓)	0.0411 (↓)	0.6483 (↓)
	LFW	0.1793 (↓)	0.5243 (↓)	0.7050 (↓)	0.1727 (↓)	0.5347 (↓)	0.6920 (↓)	0.1863 (↓)	0.5210 (↓)	0.6910 (↓)
Age	AgeDB30	0.0210 (↓)	0.0283 (↓)	0.1320 (↓)	0.0373 (↓)	0.0390 (↓)	0.1457 (↓)	0.0213 (↓)	0.0327 (↓)	0.1243 (↓)
	CALFW	0.0477 (↓)	0.0977 (↓)	0.3130 (↓)	0.0653 (↓)	0.1137 (↓)	0.3323 (↓)	0.0500 (↓)	0.1073 (↓)	0.3047 (↓)
Pose	CFP-FP	0.0020 (↓)	0.0034 (↓)	0.1323 (↓)	0.0020 (↓)	0.0043 (↓)	0.1346 (↓)	0.0017 (↓)	0.0031 (↓)	0.1146 (↓)
	CPLFW	0.0170 (↓)	0.0453 (↓)	0.1567 (↓)	0.0207 (↓)	0.0503 (↓)	0.1607 (↓)	0.0213 (↓)	0.0457 (↓)	0.1683 (↓)
Hard	VGGFace2	0.0040 (↓)	0.0508 (↓)	0.2036 (↓)	0.0052 (↓)	0.0544 (↓)	0.2160 (↓)	0.0044 (↓)	0.0488 (↓)	0.1996 (↓)
	XQLFW	0.0003 (↑)	0.0003 (↓)	0.0080 (↓)	0.0003 (↓)	0.0007 (↓)	0.0110 (↑)	0.0003 (↓)	0.0013 (↑)	0.0123 (↑)

Table 15: TAR@FAR after fine-tuning the model, with all the layers frozen except the last two, on QMUL-SurvFace.

		m=0.5			m=0.4			m=0.3		
Benchmarks		1e-4	1e-3	1e-2	1e-4	1e-3	1e-2	1e-4	1e-3	1e-2
Frontal	CFP-FF	0.0734 (↓)	0.0794 (↓)	0.8874 (↓)	0.0743 (↓)	0.0817 (↓)	0.9245 (↓)	0.0734 (↓)	0.0794 (↓)	0.9009 (↓)
	LFW	0.8316 (↓)	0.8880 (↓)	0.9516 (↓)	0.8516 (↓)	0.8923 (↓)	0.9503 (↓)	0.8417 (↓)	0.8840 (↓)	0.9517 (↓)
Age	AgeDB30	0.1307 (↓)	0.1707 (↓)	0.4283 (↓)	0.2260 (↓)	0.2390 (↓)	0.4747 (↓)	0.1303 (↓)	0.1893 (↓)	0.4237 (↓)
	CALFW	0.3793 (↓)	0.4287 (↓)	0.6457 (↓)	0.4913 (↓)	0.5370 (↓)	0.6933 (↓)	0.4120 (↓)	0.4683 (↓)	0.6550 (↓)
Pose	CFP-FP	0.0137 (↓)	0.0300 (↓)	0.4074 (↓)	0.0194 (↓)	0.0251 (↓)	0.4469 (↓)	0.0120 (↓)	0.0217 (↓)	0.4206 (↓)
	CPLFW	0.1050 (↑)	0.2400 (↑)	0.3673 (↓)	0.0830 (↑)	0.2007 (↑)	0.3977 (↓)	0.0713 (↑)	0.1890 (↑)	0.3780 (↓)
Hard	VGGFace2	0.0192 (↓)	0.2220 (↓)	0.4572 (↓)	0.0192 (↓)	0.2620 (↓)	0.5116 (↓)	0.0152 (↓)	0.2328 (↓)	0.4632 (↓)
	XQLFW	0.0000 (↓)	0.0000 (↓)	0.0047 (↑)	0.0000 (↓)	0.0000 (↓)	0.0057 (↑)	0.0000 (↓)	0.0000 (↓)	0.0040 (↓)

Table 16: TAR@FAR after fine-tuning the model, with all the layers frozen except the last two, on DigiFace-1M.