The main purpose of the following chapter is to select a superior approach to TrustID's facial recognition solution that encapsulates an appropriate trade-off between performance and computational cost. To achieve that, first the pre-trained models are benchmarked. These results, combined with the models' complexity and resources cost related specifications, will allow the choice of the most adequate model to be fine-tuned. After training, the tests are repeated in order to evaluate if the model improved or not, and a final choice for the main objective is made.

0.1 Models' specifications

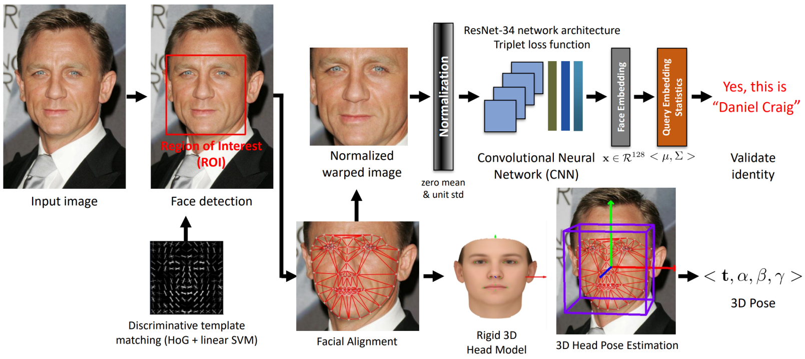
Input image
Region of Interest (ROI)
Face detection
Discriminative template matching (HoG + linear SVM)
Normalized warped image
Normalization
zero mean & unit std
ResNet-34 network architecture
Triplet loss function
Convolutional Neural Network (CNN)
Face Embedding
Query Embedding Statistics
$\mathbf{x} \in \mathcal{R}^{128}$ $< \mu, \Sigma >$
Yes, this is "Daniel Craig"
Validate identity
Facial Alignment
Rigid 3D Head Model
3D Head Pose Estimation
$< \mathbf{t}, \alpha, \beta, \gamma >$
3D Pose

Figure 1: TrustID's Architecture. Image from the case study's paper [2]

Here we compare the following four models against the TrustID Resnet-34 based solution Figure 1, which comprises 29 layers, an average inference time of 7.28 seconds (computed by averaging the inference time on the entire dataset for all the benchmarks) and has been trained on the VGG Face and Facescrub ensemble dataset using triplet loss. Relevant specifications to assist the identification of a superior model are presented in Table 1.

Table 1: Model's characteristics. "# Parameter" 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 of all images from all the benchmark datasets, "Loss" is the loss function used to train the model, and "Dataset" are the training images.

	# 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

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 should 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 pre-trained models, the mean accuracy is presented on the following table.

Table 2: Model's face verification accuracy.

Datasets	*Frontal group*		*Age group*		*Pose group*		*Hard group*	
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

From Table 2, we can see that iResnet-18 and iResnet-SE-50 have comparable performance. 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 does not, but only by a very small margin. Additionally, concerning the datasets where iResnet-SE-50 exhibited slightly lower performance, the accuracy scores are also very close. Specifically, between the two models, CFP-FF, LFW, CALFW, CFP-FP, CPLFW and VGGFace2 are all 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 did not 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 without a compromising accuracy for student monitoring. Regarding the results from the extremely hard XQLFW, they are exceedingly low, approaching 0.5 or worse, for almost all the methods. This suggests that the model is producing outputs that resemble random guesses, which is exactly what occurs with FaceNet. The only exception is TrustID, which can be probably justified by the method of training. By resizing smaller training images to 150×150 there is a degradation in quality that leads to a model more prepared to handle these situations.

0.2.2 ROC Curves

The previous table indicates that the three methods with superior performance, 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 most appropriate model to be fine-tuned, the TAR values are computed 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 prime models can be determined since those are able to correctly identify more genuine matches while keeping the incorrect matches low.

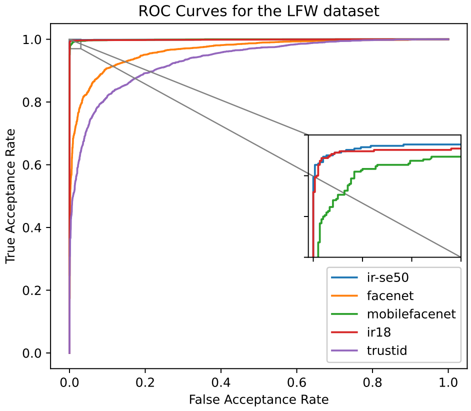

Figure 2: ROC Curves for the LFW benchmark from the Fronta

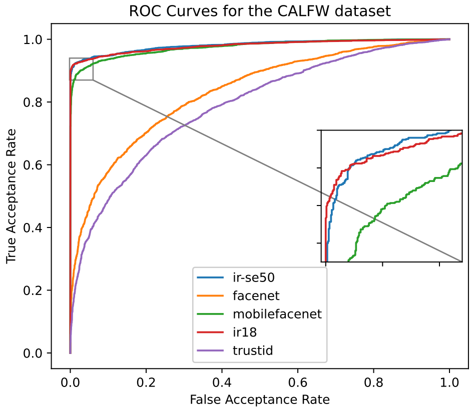

Figure 3: ROC Curves for the CALFW benchmark from the A

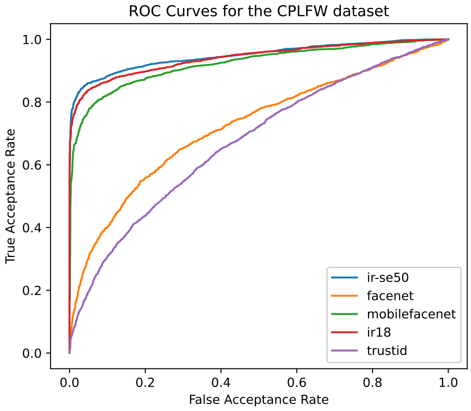

Figure 4: ROC Curves for the CPLFW benchmark from the Pose group.

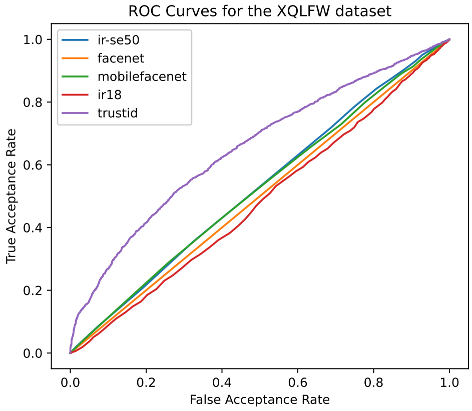

Figure 5: ROC Curves for the XQLFW benchmark from the Hard group.

The selected 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 5, 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. Please refer to Appendix E for the remaining from each benchmark group

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

		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

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

For a low FAR value of $1e-4$, the threshold is more firm and leaves less margin for identifying wrong matches as true identities, all models fail on more demanding datasets, but for less intricate ones, iResnet-SE-50 achieves suitable performance on LFW and CALFW, and MobileFaceNet on LFW. Reducing the strictness and increasing the FAR to $1e-3$, leads to 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 XQLFW dataset.

0.2.3 DET Curves

To finalize the selection of the most appropriate 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, i.e, 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, minimizing both the amount of impostors matched as true identities (FAR) and true identities classified as impostors (FRR).

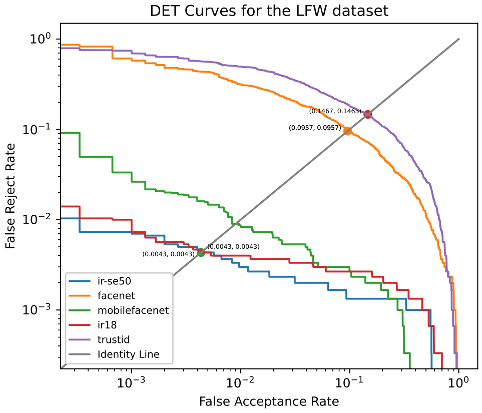

Figure 6: DET Curves for the LFW benchmark from the Frontal group.

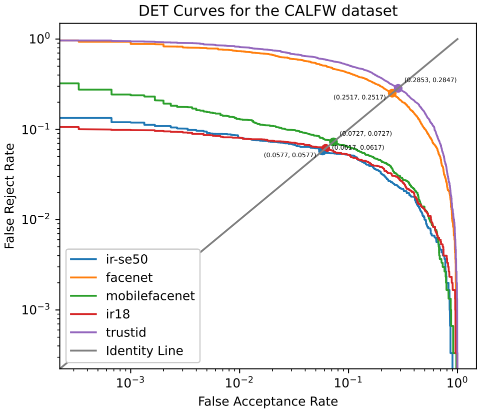

Figure 7: DET Curves for the CALFW benchmark from the Age group.

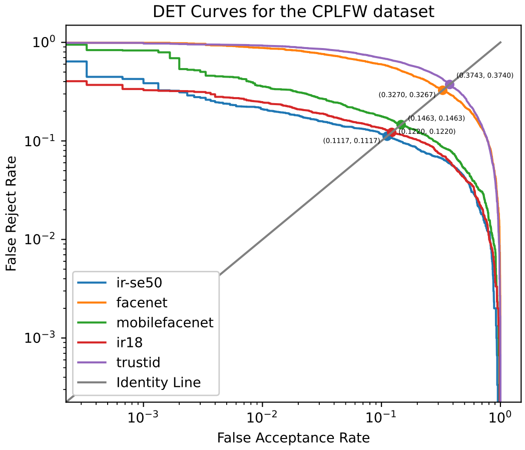

Figure 8: DET Curves for the CPLFW benchmark from the Pose group.

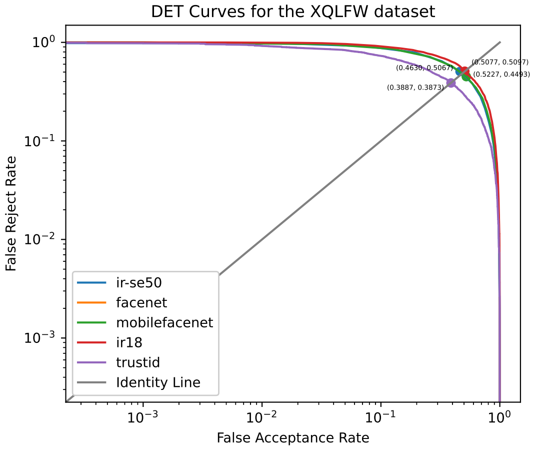

Figure 9: DET Curves for the XQLFW benchmark from the Hard group.

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

	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

As a final analysis, 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 more complex model, has the lower scores, with iResnet-18 a close second and MobileFaceNet as the third. Please refer to Appendix F for the remaining from each benchmark group

0.3 Number of trainable parameters per unit of accuracy

This metric is obtained by computing the division of the number of trainable parameters by the accuracy values, and it will serve as a complement to the metrics that evaluate resource's cost, i.e, the amount of mult-adds, inference time, and the number of trainable parameters and layers. Because the previous tests consecutively place iResnet-SE-50, iResnet-18 and MobileFaceNet as the more capable methods, they will be the ones studied in Table 5.

Table 5: Number of trainable parameters per unit of accuracy for the three highest achieving models.

Datasets / Models	Frontal		Age		Pose		Hard	
	CFP-FF	LFW	AgeDB30	CALFW	CFP-FP	CPLFW	VGGFace2	XQLFW
MobileFaceNet	1.216×10^6	1.211×10^6	1.290×10^6	1.282×10^6	1.340×10^6	1.389×10^6	1.327×10^6	2.371×10^6
iResnet-18	2.412×10^7	2.412×10^7	2.470×10^7	2.514×10^7	2.552×10^7	2.687×10^7	2.612×10^7	4.861×10^7
iResnet-SE-50	4.398×10^7	4.400×10^7	4.728×10^7	4.590×10^7	4.631×10^7	4.841×10^7	4.661×10^7	8.526×10^7

Table 2 emphasizes that the three methods attain comparable accuracy values, a crucial factor for model comparison with this metric. In this context, a lower NPUA value is preferred, as it indicates that a model can achieve the same accuracy while utilizing fewer trainable parameters. Finally, we can conclude that MobileFaceNet is the more computational cost-efficient, presenting NPUA scores lower than iResnet-18 and iResnet-SE-50 by, at least, an order of magnitude.

0.3.1 Discussion

At this stage, we are capable of assuring, based on the previous tests, that one objective is achieved and more adequate solutions to TrustID's facial verification framework were found. The three top performing methods 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.

Considering the benchmarks' results, the much lower amount of trainable parameters, mult-adds, inference time and NPUA, MobileFaceNet is the best balance between performance and computational cost. As discussed in ??, due to the monitoring context, achieving perfect accuracy is not the main concern. This is due to the fact that, as the monitoring occurs over a time span it allows the system to perform more face verifications, compensating for lower accuracy values. Furthermore, although it is true that MobileFacenet's accuracy is high in the pose group and VGGFace2, there is 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, 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.

0.4 Training Details

By leveraging Optuna's hyperparameter searching capabilities[1], we observe 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 incorrect evaluation values. Following the original ArcFace work [1], the optimizer of choice is SGD with momentum 0.9 and weight decay $5e-4$, scale $s = 64$ and margin $m = 0.5$. Manual early stopping is performed to ensure the most favorable achievable results, by evaluating on XQLFW and CPLFW during training and saving the model's training checkpoint with the highest performing accuracy. These validation datasets are selected accordingly to the areas where the models showed potential for improvement based on the benchmarks performed.

0.5 Training Results

Fine-tuning all the layers

The first approach is to update the whole network, since both QMUL-SurvFace and DigiFace-1M are large datasets with a high number of images and identities, there are less chances of quickly overfitting to the training data. Table 6 and Table 7 summarize the accuracy results for QMUL-SurvFace and DigiFace-1M, respectively.

Table 6: MobileFaceNet accuracy scores before and after fine-tuning the whole network on QMUL-SurvFace with different ArcFace margins.

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

[1] https://optuna.org/

According to Table 6, fine-tuning with QMUL-SurvFace improves, as intended, 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. Hence, in an effort to improve the results, the margin is reduced in order to generate a less penalizing training with a smaller distance between classes. The results from these settings are shown in Table 6, and when fine-tuned with QMUL-SurvFace, it occurs the same behavior as in $m = 0.5$. The XQLFW results are also better than the original pre-trained model, increasing 5.77% for $m = 0.4$ and 3.00% for $m = 0.3$, while the other scores worsen.

Table 7: MobileFaceNet accuracy scores before and after fine-tuning the whole network on DigiFace-1M with different ArcFace margins.

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

In the case of DigiFace-1M in Table 7, no discernible improvements were observed, and reducing the margin size did not yield any positive changes. Instead, the model appears 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 inferior results in the frontal and age groups for $m = 0.4$ and $m = 0.3$, the pose group is superior for both margins and the hard group is mixed, where VGGFace2 has lower 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, aside from the XQLFW when tuned with QMUL-SurvFace.

To achieve a more profound understanding of how the model reacts to the data, Table 8 and Table 9 present the TAR at very low FAR.

Table 8: TAR@FAR after fine-tuning the model with 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	0.0140 (↓)	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 (↓)

The reduced margins $m = 0.4$ and $m = 0.3$ produce lower TAR scores at any FAR, including the very relevant frontal group. However, when $m = 0.5$, there are some improvements. At $FAR = 1e-3$ and $FAR = 1e-2$, the pose group suffers a significant improvement and XQLFW increases slightly.

Table 9: TAR@FAR after fine-tuning the model with DigiFace-1M.

		$m = 0.5$			$m = 0.4$			$m = 0.3$		
Benchmarks		$1e-4$	$1e-3$	$1e-2$	$1e-4$	$1e-3$	$1e-2$	$1e-4$	$1e-3$	$1e-2$
Frontal	CFP-FF	0.0020 (↓)	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 (↑)

Once again, the model trained with DigiFace-1M, does not perform and loses discriminative power at any margin and FAR value with the exception of a few outliers values that increase minimally, which can be seen as a mere fluctuation.

In the context of general CNN architectures, it is 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 followed: 1) freeze the first 5 layers and 2) train only the 2 final layers. Additionally, based on the previous experiences, different ArcFace margins are also studied.

Fine-tuning with 5 initial frozen layers

Table 10 and Table 11 present the accuracy scores after fine-tuning MobileFaceNet without updating the first five layers' weights. By freezing only 5 layers, the model still keeps its ability to learn more complex information associated with final stages of the network.

Table 10: MobileFaceNet accuracy scores before and after 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	0.7821 (↓)	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 11: MobileFaceNet accuracy scores before and after fine-tuning the network, with the first five layers frozen, on DigiFace-1M 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

As expected, the weights are less updated, hence the majority of the accuracy scores are higher than when training the whole network. Also, for QMUL-SurvFace (Table 10), if $m = 0.3$ there is even an improvement in the XQLFW benchmark. Finally, DigiFace-1M (Table 11) at any m value, shows no positive developments against the pre-trained model.

Table 12: 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	0.0117 (↓)	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 13: TAR@FAR after fine-tuning the model, with the first five layers frozen, on DigiFace-1M.

		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 (↑)

Finally, Table 12 and Table 13 highlight that training with less layers does not improve the performance at any FAR or margin. Moreover, the previous enhancements verified when the complete model is trained with QMUL-SurvFace with $m = 0.5$, and $FAR = 1e - 3$ and $FAR = 1e - 2$ are lost with this configuration.

Fine-tuning the last 2 layers

To finalize, we conducted additional tests by training only the last two layers (one convolutional and linear). This allows us to investigate in which direction the model evolves in terms of results when even less stages are trained.

Table 14: MobileFaceNet accuracy scores 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	0.8751 (↓)	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 15: MobileFaceNet accuracy scores before and after fine-tuning the network, with all the layers frozen aside the last two, on DigiFace-1M 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 and Table 15 follows the pattern seen in the previous experiment. On one hand, training less layers produces results closer to the pre-trained model, on the other, the network does not improve in any meaningful way.

Table 16: 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	0.038 (↓)	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 17: TAR@FAR after fine-tuning the model, with all the layers frozen except the last two, on DigiFace-1M.

		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 (↓)

To conclude, the TAR values continue to be worse than the original model, for both QMUL-SurvFace (Table 16) and DigiFace-1M (Table 17), with the usual outliers values that are slightly higher but do not highlight any pattern of improvement.

0.5.1 Discussion

Taking into consideration the previous experiments, it can be inferred that the most adequate solution is the pre-trained model. Further training on any of the selected datasets does not improve the overall performance. Although, fine-tuning the whole network with QMUL-SurvFace leads to higher accuracy on the XQLFW benchmark, the scores for the other tests are worse. Moreover, when $m = 0.5$ the model is more discriminative for $FAR = 1e{-}3$ and $1e{-}2$ on the pose group and XQLFW, but once again, on the other benchmarks the model becomes less discriminative. Since the intended application is to monitor students, it is highly required for the model to adapt to any quality and pose, that is, profile and frontal. Therefore, improvement on one benchmark at a cost of having negative impact on the others is less than idealized. The initial pre-trained model outperforms TrustID's method, with low computational cost and respectable scores on all the tests and benchmarks. Despite the fact that it did not respond well to the fine-tuning on the selected datasets, it still is the most competent approach to the student monitoring problem as it performs appropriately considering the possible challenges of pose, illumination, expressions, etc. while maintaining a light resource utilization[2].

[2] A video demonstration of the chosen method on a real-time webcam feed is available in the following repository:`https://github.com/davidmcarreira/dfrosi-demo`