Improving Training and Inference of Face Recognition Models via Random Temperature Scaling

Lei Shang[1,*], **Mouxiao Huang**[2,3,*], **Wu Shi**[2,†], **Yuchen Liu**[1], **Yang Liu**[1], **Fei Wang**[1], **Baigui Sun**[1,‡], **Xuansong Xie**[1], **Yu Qiao**[2,4]

[1] Alibaba Group
[2] The Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences
[4] Shanghai Artificial Intelligence Laboratory
{sl172005, yuchen.lyc, ly261666, steven.wf, baigui.sbg, xingtong.xxs}@alibaba-inc.com
{mx.huang, wu.shi, yu.qiao}@siat.ac.cn

Abstract

Data uncertainty is commonly observed in the images for face recognition (FR). However, deep learning algorithms often make predictions with high confidence even for uncertain or irrelevant inputs. Intuitively, FR algorithms can benefit from both the estimation of uncertainty and the detection of out-of-distribution (OOD) samples. Taking a probabilistic view of the current classification model, the temperature scalar is exactly the scale of uncertainty noise implicitly added in the softmax function. Meanwhile, the uncertainty of images in a dataset should follow a prior distribution. Based on the observation, a unified framework for uncertainty modeling and FR, Random Temperature Scaling (RTS), is proposed to learn a reliable FR algorithm. The benefits of RTS are two-fold. (1) In the training phase, it can adjust the learning strength of clean and noisy samples for stability and accuracy. (2) In the test phase, it can provide a score of confidence to detect uncertain, low-quality and even OOD samples, without training on extra labels. Extensive experiments on FR benchmarks demonstrate that the magnitude of variance in RTS, which serves as an OOD detection metric, is closely related to the uncertainty of the input image. RTS can achieve top performance on both the FR and OOD detection tasks. Moreover, the model trained with RTS can perform robustly on datasets with noise. The proposed module is light-weight and only adds negligible computation cost to the model.

1 Introduction

Recently, deep neural networks have achieved great success in face recognition. State-of-the-art face recognition methods generally map each face image to an identity-related representation in the latent space (**?????**). Ideally, the representations of the same person are embedded in a compact cluster. However, when the face image contains uncertainty and noise, the learned representation may be unreliable and tend to cause mistakes.

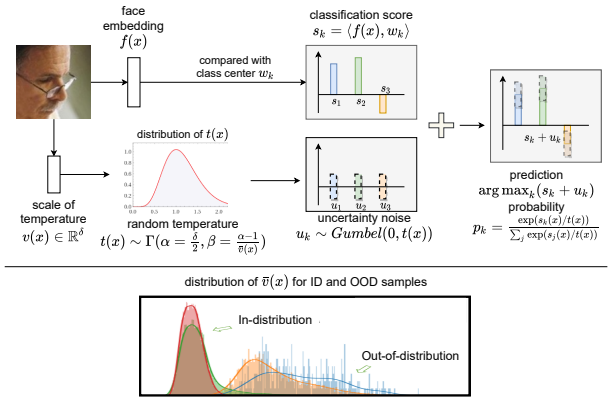


Figure 1: Random Temperature Scaling models the data uncertainty via the scale of temperature (top). We find that the learned scale, $v(\boldsymbol{x})$, is a good metric for out-of-distribution detection (bottom).

It is well known that deep learning models can achieve high accuracy in many classification tasks given a relatively clean and closed dataset. However, in the real world, the data are captured in more unconstrained scenarios (**??**). Data uncertainty commonly present in the images for face recognition. For example, in surveillance videos (**???**), the image can be corrupted by motion blur, focus blur, and lighting conditions. When key facial landmarks are occluded due to pose, hair, or accessories, it is not confident to extract accurate representations for matching. In the step of face detection, partial face or non-face images can be proposed for recognition. The results for these data can be unpredictable. It is essential to estimate the uncertainty of data and detect OOD samples for safety-critical algorithms like face recognition.

Probabilistic face embeddings (PFE) (**?**) is the first to consider data uncertainty in face recognition. Instead of a fixed representation, the face is embedded as a Gaussian distribution whose mean is the original representation and variance is estimated by another model. A new similarity metric,

[*] These authors contributed equally.
[†] Corresponding author.
[‡] Project lead.

mutual likelihood score (MLS), is proposed to measure the matching likelihood between two images. The main limit of PFE is that it can not learn the representation at the same time. Data uncertainty learning (DUL) (**?**) also models the face representation as a Gaussian distribution. It uses the re-parameterization trick to simultaneously learn the feature (mean) and uncertainty (variance). The learned features can be directly used for conventional similarity metrics with better intra-class and inter-class properties. The learned uncertainty affects the model by adaptively reducing the adverse effects of noisy images.

Current state-of-the-art OOD methods are generally designed for classification tasks. ODIN (**?**) provides two strategies, temperature scaling and input preprocessing, for OOD detection, based on a trained neural network classifier. It finds that the maximum class probability (called *softmax score*) is an effective score for detecting OOD data. Despite its effectiveness, ODIN requires OOD data to tune hyperparameters. This requirement could prevent the tuned hyperparameter to generalize to other datasets (**?**). GODIN (**?**) extends the setting of ODIN, and proposes two strategies, data-dependent temperature scaling and auto-tuned input preprocessing, for learning without OOD data. Techapanurak *et al*. (**?**) propose a hyperparameter-free method based on softmax of scaled cosine similarity. The proposed method shows at least comparable performance to the state-of-the-art methods on the conventional test, and outperforms the existing methods on the recently proposed test (**?**). Although these methods are not designed for face recognition models, it is worth conducting experiments to evaluate the performance on OOD detection when learning face representations at the same time.

We propose an effective method, Random Temperature Scaling (RTS), to learn a reliable face recognition model. First we take a probabilistic view of the classification method using softmax with temperature scaling, and find that the temperature scalar is exactly the scale of uncertainty noise implicitly used in softmax. The larger the temperature is, the larger the variance of uncertainty becomes. Instead of a fixed scalar, RTS models the stochastic distribution of temperature which depends on the input data. RTS acts like a Bayesian network (**??**) and has a regularization effect on the confidence. Both the face representation and the temperature can be learned simultaneously. Besides, RTS adds only negligible computation cost to the original face recognition model. Our contributions include: (1) We reveal the connection between temperature scaling and uncertainty in classification and propose Random Temperature Scaling (RTS) to learn reliable face recognition models. (2) In the training phase, RTS can adjust the learning strength of clean and noisy samples for stability and accuracy. (3) In the test phase, RTS can provide an uncertainty score to detect uncertain, low-quality and even OOD samples, without training on extra labels.

2 Related Work

Uncertainty in face recognition. Uncertainty in deep learning models is getting more attention these years. How uncertainty affects the performance of deep neural networks is investigated in many works (**???**). The pioneering work (**?**) introduces two types of uncertainty for deep neural networks via Bayesian modeling. Data uncertainty captures noise in the data distribution, and model uncertainty accounts for uncertainty in the model parameters. Model uncertainty can be explained away given enough data. In the real world, data uncertainty commonly presents in the images for face recognition. Modeling data uncertainty is essential for reliable deep learning models. PFE (**?**) is the first work to consider data uncertainty in face recognition task. In PFE, each face is embedded as a Gaussian distribution whose mean is the original representation and variance is estimated by another model. The main limit of PFE is that it relies on the pre-trained face representations. DUL (**?**) is the first work to train the face representation and uncertainty at the same time.

Out-of-distribution detection. It is known that deep learning algorithms often make high confidence predictions even for unrecognizable or irrelevant objects (**??**). Besides capturing such uncertainty in classifiers, detecting the OOD sample is a more direct approach to that problem. ODIN (**?**) finds that the maximum class probability (called *softmax score*) is an effective score for detecting OOD data, and provides two strategies, temperature scaling and input pre-processing, for OOD detection. Despite its effectiveness, ODIN requires OOD data to tune hyperparameters. GODIN (**?**) extends the setting of ODIN, and proposes two corresponding strategies for learning without OOD data. Techapanurak *et al*. (**?**) propose a hyperparameter-free method based on softmax of scaled cosine similarity, and experiments show its competitive performance in many OOD tests.

Temperature scaling. Temperature scaling is a widely used technique for confidence calibration (**??**), knowledge distillation (**?**), and adaptive classification objectives (**?**). In knowledge distillation (**?**), the temperature scalar is a global value tuned for each task and each model to soften the prediction. In (**?**), the temperature is dynamically adjusted to learn a better representation. In confidence calibration (**?**), the temperature is learned depending on the input data. We find the temperature scalar is connected to the scale of uncertainty in the classifier.

3 Methodology

3.1 Preliminaries

A face recognition algorithm generally maps an image $\boldsymbol{x} \in \mathbb{R}^{h\times w\times 3}$ to an identity-related representation $\boldsymbol{y} = f(\boldsymbol{x}) \in \mathbb{R}^{d_y}$. The predicted probability is defined by

$$p_k(\boldsymbol{x}) = \frac{\exp(s_k(\boldsymbol{x}))}{\sum_j \exp(s_j(\boldsymbol{x}))}, \quad s_j(\boldsymbol{x}) = \langle f(\boldsymbol{x}), \boldsymbol{w}_j \rangle, \quad (1)$$

where $\boldsymbol{w}_j$ is called the class center for identity $j \in \{1, \ldots, C\}$ and C is the number of different classes. Given a set of images $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ and their labels $\{c(\boldsymbol{x}_1), \ldots, c(\boldsymbol{x}_N)\}$, the classifier is trained by minimizing

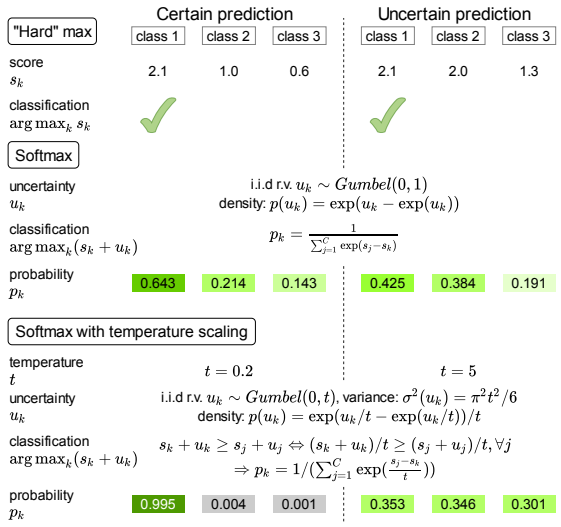


Figure 2: A probabilistic view of softmax and temperature scaling. The standard softmax function models the uncertainty of prediction by implicit Gumbel random variables. In temperature scaling, the temperature corresponds to the scale of uncertainty noise added to the classification score.

the Cross Entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(p_{c(\boldsymbol{x}_i)}(\boldsymbol{x}_i)\right). \tag{2}$$

To improve the discriminative ability of the learned features, various margin-enhanced losses (**??**) are proposed. For example, the ArcFace function $s_k = ArcFace(\boldsymbol{y}, c, \{\boldsymbol{w}_j\}, k; \gamma, \theta)$ is defined by:

$$s_k = \begin{cases} \gamma \cdot \cos(\cos^{-1}(\langle \boldsymbol{y}, \boldsymbol{w}_k \rangle) + m), & k = c \\ \gamma \cdot \langle \boldsymbol{y}, \boldsymbol{w}_k \rangle, & k \neq c \end{cases}, \tag{3}$$

where c is the corresponding label of $\boldsymbol{y}$, γ is the scaling factor and m is the additive arc margin. Substituting Eqn. 3 into Eqn. 1, we can derive the predicted probability by ArcFace.

3.2 Random Temperature Scaling

In the following, we elaborate on the principles and details of the proposed method, Random Temperature Scaling (RTS), for face recognition with uncertainty.

A probabilistic view of softmax and temperature scaling. As shown in Fig. 2, a typical classification model predicts the score, s_k, for each possible class. To make a decision, the model chooses the class with the maximum score, which can be regarded as a "Hard" max method (top). However, this method is deterministic and cannot indicates the confidence of results. Intuitively, the prediction made from scores (2.1, 1.0, 0.6) should be more confident than that from (2.1, 2.0, 1.3), though the two cases result in the same prediction. To consider the confidence, a noise, u_k, can be added to the score, and the model predicts from the score $s_k + u_k$ with uncertainty (middle). Assuming that u_k's are i.i.d. Gumbel random variables, the probability of class k is given by:

$$\Pr(s_k + u_k \geq s_j + u_j, \forall j) = \frac{1}{\sum_{j=1}^{C} \exp(s_j - s_k)}. \tag{4}$$

The scale of uncertainty noise can vary by a factor t, *i.e.*, $u_k \sim Gumbel(0, t)$, to reflect the different uncertainties of inputs. Through the change of variables, we can obtain

$$\Pr(\frac{s_k + u_k}{t} \geq \frac{s_j + u_j}{t}, \forall j) = \frac{1}{\sum_{j=1}^{C} \exp(\frac{s_j - s_k}{t})}. \tag{5}$$

Eqn. 5 is exactly the softmax function with temperature scaling. The temperature corresponds to the scale of uncertainty noise, u_k, added to the classification score (Fig. 2 bottom). The variance of u_k is $\pi^2 t^2/6$. The larger the temperature is, the larger the variance of uncertainty becomes.

A stochastic distribution of the temperature. We assume that the temperature in a dataset follows a prior distribution, we propose the Random Temperature Scaling (RTS) method to model the stochastic distribution of temperature. In RTS, the temperature, $t(x)$, is a random variable whose scale depends on the input data. The algorithm is described in Alg. 1. Specifically, f is the head for face representation and g is the head for log scale of the temperature scalar. The temperature is a sum of δ independent Gamma variables: $r_i = v_i \cdot \epsilon_i^2 \sim \Gamma(\frac{1}{2}, \frac{1}{2v_i})$. For illustration, we let $v_i = v, \forall i$. We divide the sum by $(\delta - 2)$ to let the mode (most frequent value) of t to be equal to v. Thus, t follows $\Gamma(\alpha = \frac{\delta}{2}, \beta = \frac{\alpha - 1}{v})$. The effects of δ and v are shown in Fig. 3. The degree of freedom, δ, is a hyperparameter and the scale, $v(x)$, is a learned function of the input image.

Difference between Relaxed Softmax and RTS. Relaxed Softmax (**??**) learns to predict a fixed uncertainty level for each image, while RTS models the uncertainty level by a stochastic distribution. Relaxed Softmax tends to increase the uncertainty level at early stage and doesn't turn back the trend to result in an overly smoothed prediction. RTS has a regularization effect on the confidence. The dynamic of temperature during training is depicted in Fig. 4. The distribution gradually becomes close to its prior. This encourages the classifier to make a correct prediction, instead of simply tuning down its confidence.

Algorithm 1: Random Temperature Scaling (RTS)

Input: Image $\boldsymbol{x}$, label c, dof δ, scale γ, margin m
Parameter: $\theta_f, \theta_g, \{\boldsymbol{w}_j\}$
Output: Scores $\boldsymbol{s}$

face repr. $\boldsymbol{y} = f(\boldsymbol{x}; \theta_f)$
log scale $z = g(\boldsymbol{x}; \theta_g) \in \mathbb{R}^{\delta}$
scale $v = \exp(z) \in \mathbb{R}^{\delta}$
$\epsilon_i \sim \mathcal{N}(0,1), \forall i = 1, \ldots, \delta$
$t = \frac{1}{\delta - 2} \cdot \sum_i v_i \cdot \epsilon_i^2$
$s'_k = ArcFace(\boldsymbol{y}, c, \{\boldsymbol{w}_j\}, k; \gamma, \theta), \forall k = 1, \ldots, C$
$\boldsymbol{s} = [s'_1/t, \ldots, s'_C/t]$

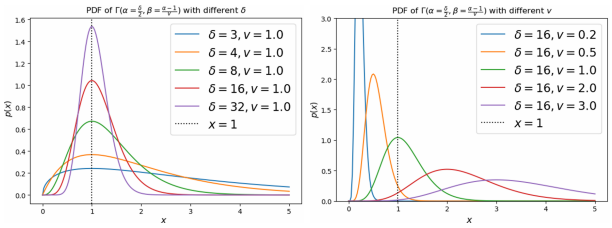


Figure 3: The distributions of t with different δ and v.

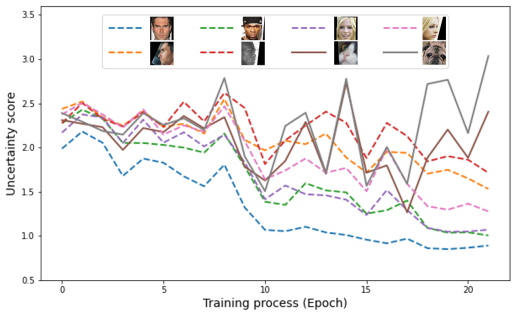


Figure 4: Variation of uncertainty score in RTS during the training process. Training data: DeepGlint.

Difference between DUL and RTS. DUL (**?**) adds explicit uncertainty noise to the feature embedding, while RTS employs the implicit noise in the softmax function to model the uncertainty. For DUL, adding randomness into the representation leads to ambiguity of learning objective (**?**), and sampling from the high dimensional distribution is not efficient for training (**?**). RTS mildly changes the confidence, instead of changing the predicted result.

3.3 Training Objectives

We train the classifier using softmax with RTS. In addition to the face feature $f(\boldsymbol{x})$, the classifier also predict a log scale, $g(\boldsymbol{x}) = \log v(\boldsymbol{x}) \in \mathbb{R}^{\delta}$, of the underlying Gamma variable. In our implementation, f and g share the same backbone neural network and use different prediction heads. The scaled logits, $\boldsymbol{s}$, are given in Alg. 1. The Cross Entropy loss, $\mathcal{L}_{CE}$, can be obtained by substituting $\boldsymbol{s}$ into Eqn. 2. Besides, we add a constraint that encourages each independent Gamma variable, r_i, to follow the prior distribution $\Gamma(\frac{1}{2}, \frac{1}{2})$:

$$\begin{aligned}\mathcal{L}_{KL} =& \frac{1}{\delta}\sum_{i=1}^{\delta} D_{KL}\left(\Gamma(\frac{1}{2}, \frac{1}{2v_i}), \Gamma(\frac{1}{2}, \frac{1}{2})\right) \\ =& \frac{1}{\delta}\sum_{i=1}^{\delta} \frac{1}{2}(v_i - \log v_i - 1)\end{aligned} \tag{6}$$

Finally, the overall loss is given by

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot \mathcal{L}_{KL}, \tag{7}$$

where λ is a hyperparameter to balance the two losses.

4 Experiments

4.1 Implementation Details

Datasets. We use CASIA-WebFace (10k IDs / 0.5M Images) (**?**) and DeepGlint (181k IDs / 6.75M Images) (**?**) as the training data for our experiments. For face recognition task, seven benchmarks including LFW (**?**), CFP (**?**) (frontal-frontal and frontal-profile protocol), AgeDB (**?**), CALFW (**?**), CPLFW (**?**), and VGG2 (**?**) are used to evaluate the performance of the baseline and our methods. For OOD detection task, we build an OOD test set consisting of human face, cat and dog images. For in-distribution data, we use LFW (**?**) and random select 15339 images from RAFDB (**?**). All the face images are not seen in the training phase. For out-of-distribution data, we collect 349 dog images and 3671 cat images. It is worth noting that all the OOD images are misclassified by a face detector as face images. Examples are shown in Fig. 4.

Training details. We use ResNet (**?**) backbone with SE (**?**) blocks as the baseline model (Backbone: ResNet64). The dimension of face embedding is 512. All the models use the same backbone. Baseline, GODIN (**?**), Relaxed Softmax (**??**), DUL (**?**), and RTS use ArcFace (**?**) as the prediction head while MagFace (**?**) and AdaFace (**?**) uses their own prediction heads. Our model (RTS) has extra output dimensions which corresponds to the log-variance $g(\boldsymbol{x})$ of the underlying Gamma variable. According to the characteristics of methods, we divide all models into two categories: non-uncertainty models (Baseline, MagFace and AdaFace) and uncertainty models (GODIN, Relaxed Softmax, DUL and RTS). We follow ArcFace (**?**) and DUL (**?**) to set experimental parameters. On CASIA-WebFace, the initial learning rate is 0.1, and is decreased to 0.01 and 0.001 after epoch 20 and 28. The training process is finished at epoch 32. On DeepGlint, models are trained for 22 epochs. The learning rate starts from 0.1 and is divided by 10 at epoch 10 and epoch 18. We use the SGD optimizer with a momentum of 0.9, weight decay of 0.0005 and batch size of 512. Besides, during the experiments, we found that the convergence of uncertainty models and variation of distributions of dataset during the training process are sensitive to the margin in ArcFace especially for Relaxed Softmax. Small margin at the beginning makes models easier to converge. Thus the margin changes linearly in the former half of the training process and remains unchanged in the latter half of the training process. And it is worth noting that this training trick has no obvious impact on the final performance (both verification benchmarks and out-of-distribution metrics). In order to more rigorously eliminate the impact of this training trick, in our comparative experiments, all models with ArcFace prediction head uses dynamic margin without particular instructions.

4.2 Training Phase

Variation of Uncertainty Score. To eliminate the effect of dynamic margin for the variation of uncertainty score during the training process of RTS, we use fixed margin in this section to illustrate that our method can adjust the learning strength of clean and noisy samples for stability

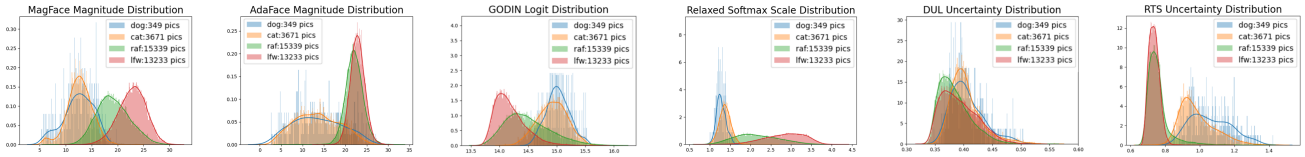


Figure 5: Distributions of magnitudes or uncertainty scores on the OOD dataset. From left to right, each subplot corresponds to the model trained with MagFace, AdaFace, GODIN, Relaxed Softmax, DUL and RTS respectively. LFW and RAFDB are considered as in-distribution data. Cat and dog are considered as out-of-distribution data. All the models are trained on DeepGlint.

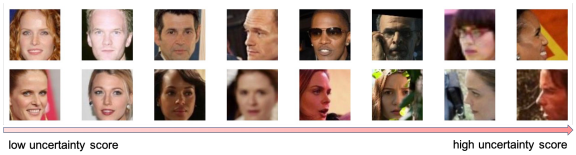


Figure 6: Images with different uncertainty scores estimated by RTS.

pct.	method	LFW	AgeDB	CPLFW	VGG2-FP
0%	baseline	99.38	93.95	89.88	93.88
	DUL	99.33	94.02	89.18	93.80
	RTS	**99.43**	**94.43**	**90.55**	**94.56**
20%	baseline	99.40	94.25	89.40	93.80
	DUL	99.23	92.82	88.88	93.36
	RTS	**99.45**	**94.42**	**90.13**	**94.09**

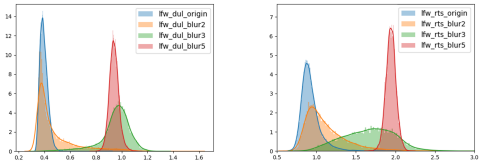
Table 1: Results of recognition with image noise. Training data: CASIA-WebFace

and accuracy. As can be seen in Fig. 4, in the early stage, all images have similar high uncertainty scores and the uncertainty score tends to increase as the prediction is inaccurate and random. As the training proceeds, the prediction for high quality data becomes more accurate and the uncertainty score decreases for the data with high confidence while increases for the data with low confidence. In the end, RTS has the ability to distinguish different quality images, giving high-quality images (frontal faces) with low uncertainty and giving low-quality images (profile faces and non-human faces) with high uncertainty.

Training with Noisy Datasets. Training on noisy datasets can significantly reduce the accuracy of deep learning models. Learning to fit bad examples indiscriminately will influence the quality of embedding for normal samples. DUL claims that it can attenuate the effect from ambiguous representations caused by poor quality samples during training. Our learned scale $v(\boldsymbol{x})$ can also balance the contributions of different samples in the learning process. The larger $v(\boldsymbol{x})$ is, the less the uncertain sample contributes to the gradient of computation. We conduct experiments training with noisy CASIA-WebFace to illustrate the robustness of our method. Specifically, we random select different proportions of images from CASIA-WebFace and add Gaussian blur to them. As shown in Table 1, RTS is robust to noisy training data.

Method	TNR @TPR90	TNR @TPR95	AUC
MagFace	79.68	63.90	97.48
AdaFace	87.33	84.04	94.19
GODIN	59.03	27.97	89.04
Relaxed Softmax	96.80	76.15	96.34
DUL	39.46	31.05	66.85
RTS	**99.60**	**98.13**	**98.38**

Table 2: Performance of different methods on OOD testset.



(a) DUL (b) RTS

Figure 7: Uncertainty scores vs. blur levels (Gaussian blur radius are 2, 3 and 5, respectively). See Appendix for distributions of other types of noise. Training data: DeepGlint.

4.3 Testing Phase

Relation Between Uncertainty and Image Quality. Given the estimated uncertainty scores $v(\boldsymbol{x})$ by RTS, we visualize the images with different levels of uncertainty in Fig. 6. For images with low uncertainty scores, the faces are generally clean and frontal. As the uncertainty level increases, the faces gradually become blurred, occluded, and corrupted by bad lighting conditions and large pose angles. Besides the data uncertainty, the score also reflects the learning difficulty for the input image.

Relation Between Uncertainty and Image Noise. To further explore the meaning of the learned uncertainty and its relation to image noise, we evaluate RTS and DUL on noisy test set. As shown in Fig. 7, we corrupt images from LFW with different Gaussian blur radius to obtain new data set LFW-blur. The distributions of uncertainty scores are shown in Fig. 7 (a - b). As the noise level increases, the predicted uncertainty scores of our RTS gradually shift to large values, while the scores of DUL change irregularly. See Appendix for distributions of other types of noise. This indicates that the scale of RTS is closely related to the intensity of noise

	Method	LFW	CFP-FF	CFP-FP	AgeDB	CALFW	CPLFW	VGG2-FP	Avg.
w/o uncertainty	Baseline	99.80	99.67	97.95	97.90	96.07	92.58	95.90	**97.12**
	MagFace	99.78	99.71	97.96	97.70	95.95	92.13	95.60	96.98
	AdaFace	99.81	**99.82**	97.87	97.98	96.07	**92.83**	94.96	97.05
w/ uncertainty	PFE (original)	**99.82**	-	93.34	-	-	-	-	-
	GODIN	99.80	99.70	98.08	**98.15**	95.98	91.85	95.64	97.03
	Relaxed Softmax	99.68	99.71	97.83	97.97	95.88	92.32	95.50	96.98
	DUL	99.78	99.72	97.92	97.95	**96.15**	92.66	95.22	97.06
	RTS	99.77	99.74	**98.09**	97.98	95.90	92.32	**96.02**	**97.12**

Table 3: Results of recognition trained on DeepGlint (except PFE). The results are all comparably high enough.

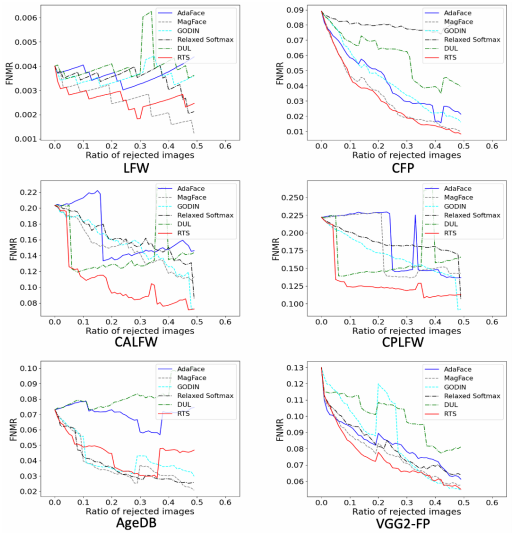


Figure 8: Face recognition performance with rejecting low-quality images. The curves show the effectiveness of rejecting low-quality face images in terms of false non-match rate (FNMR) at false match rate (FMR) threshold of 0.001. Training data: DeepGlint.

and the quality of image.

Face Recognition with Rejecting Low-quality Images. Fig. 8 shows the error-versus-reject curves of rejecting different quality face images in terms of false-non-match rate (FNMR). In order to control variables to illustrate the performance of each model on rejection task, we first obtain image features which are used for calculating FNMR from RTS. For non-uncertainty model (MagFace, AdaFace), we use magnitudes of image features obtained from MagFace and AdaFace to reject low-quality images. And for uncertainty models, we use the proposed OOD score in each model (Details can be seen in 4.4) to reject poor quality samples. Dropping low-quality faces can benefit face recognition performance significantly. As shown in Fig.8, RTS achieves the best FNMR on different ratio of rejected images in all benchmarks except LFW and AgeDB. RTS performs best in the former 20% of LFW-benchmark and former 8% of AgeDB-benchmark (only rejecting a small amount of samples), and has high performance uniformity for various testsets. It is

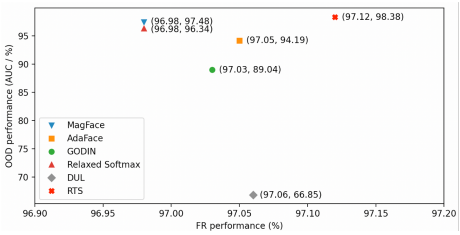


Figure 9: Balance between FR and OOD performance. Training data: DeepGlint.

noteworthy that in real application scenarios of face recognition, the proportion of low-quality images is usually small and only a small number of samples will be rejected for recognition. Considering this, our method can achieve best performance in real application. RTS is able to distinguish samples with low-quality and give the corresponding sample a reasonable uncertainty score. Compared with other methods, RTS can stably complete rejection task to improve the performance of face recognition.

4.4 Out-of-Distribution Detection

Uncertainty Scores and Evaluation Metrics. To evaluate the performance of uncertainty models on OOD detection task reasonably, we use the proposed score in each model. For GODIN and Relaxed Softmax, the proposed score is the multiplier or denominator to the logit. For DUL, the proposed score is the harmonic mean of estimated variance. We have shown that the estimated value of $v(\boldsymbol{x})$ in RTS is related to the scale of temperature, and thus reflects the scale of uncertainty in the classification model. We choose $v(\boldsymbol{x})$ as the uncertainty score of our method to detect OOD samples. For non-uncertainty model (MagFace, AdaFace), we use magnitudes of image features obtained from MagFace and AdaFace to detect OOD samples. Following literature (**?**), we use true negative rate at $\{90\%, 95\%\}$ true positive rate (TNR@TPR90, TNR@TPR95) and the area under the receiver operating characteristic curve (AUC) to evaluate OOD detection performance.

OOD Detection Performance. Table 2 shows that RTS outperforms the other methods on OOD detection task. MagFace, AdaFace and Relaxed Softmax perform well. GODIN has a relatively good OOD result. While DUL cannot detect out-of-distribution data accurately. Besides, the distributions

λ	LFW	CFP-FP	AgeDB	CPLFW	VGG2-FP
0.1	98.43	93.53	87.63	84.12	91.62
1	99.23	95.57	91.97	87.03	93.28
10	**99.43**	**97.50**	**94.43**	**90.55**	**94.56**

Table 4: Results of our models trained with different weights for KL divergence (δ is 16). Training data: CASIA-WebFace.

δ	LFW	CFP-FP	AgeDB	CPLFW	VGG2-FP
8	99.30	97.11	93.75	89.33	94.38
16	**99.43**	**97.50**	**94.43**	**90.55**	**94.56**
32	99.33	97.04	93.47	89.62	93.70

Table 5: Results of our models trained with different values for degree of freedom (λ is 10). Training data: CASIA-WebFace.

of different models are shown in Fig. 5. We can see that our proposed RTS can better discriminate the in-distribution and out-of-distribution data. Both quantitative and distributions results indicate that RTS is an effective technique to model uncertainty and complete OOD detection task.

4.5 Face Recognition Performance

Face verification accuracy on benchmarks. We compare RTS with the state-of-the-art uncertainty methods including PFE, GODIN, DUL and Relaxed Softmax, and related non-uncertainty methods including Baseline (ArcFace), MagFace and AdaFace. The results of recognition is shown in Table 3. We can see that the recognition performance of RTS is comparable with the state-of-the-art methods on all test sets. This indicates that, besides the ability to reveal uncertainty of images and detect out-of-distribution data, the model trained with RTS can achieve competitive performances in face recognition task.

Balance Between FR and OOD Performance. Fig. 9 shows the performance between face recognition (average verification accuracy of all benchmarks) and out-of-distribution detection (AUC) of non-uncertainty and uncertainty models. The comprehensive performances of MagFace, AdaFace and Relaxed Softmax are both close to that of RTS. While MagFace has a great many hyper parameters needed to be adjusted manually and is difficult to reproduce good enough results. Besides, the convergence of Relaxed Softmax is very sensitive to the margin in its prediction head. In comparison, RTS has less hyper parameters and is easier to converge. Our method achieves the best performance in OOD detection task and comparably high enough face verification accuracy, demonstrating that RTS is a unified framework for uncertainty estimation and face recognition.

4.6 Ablation Study

Effects of KL divergence. The KL divergence loss works as a regularization term to prevent the uncertainty scale from growing infinitely. When the weight $\lambda < 0.1$, the model have difficulty in converging, and the performance also deteriorates at last. For large $\lambda(> 10)$, the model tends to predict nearly constant variance $v(\boldsymbol{x})$, which has little effects in modeling data uncertainty. We conduct experiments on models with $\lambda \in \{0.1, 1, 10\}$. The results are shown in Table 4. Through experiments, we find that the model achieves the best performance when $\lambda = 10$. Thus, we set $\lambda = 10$ for our RTS model.

Effects of degree of freedom. We study the effects of degree of freedom δ, which is a hyperparameter of RTS. Intuitively, δ determines the shape of density of random temperature, t. The results are shown in Table 5. From experimental results, we can see that RTS achieves the best performance when $\delta = 16$. Thus, we set $\delta = 16$ for our RTS model.

5 Conclusion

In this paper, we first analysis the connection between temperature scaling and uncertainty modeling in the classification model. Taking a probabilistic view, the temperature scalar is exactly the scale of uncertainty noise implicitly added in the softmax function. Based on this observation, a unified framework, Random Temperature Scaling (RTS), is proposed for uncertainty estimation and face recognition by modeling the uncertainty level by a stochastic distribution.

Experiments show that RTS can adjust the learning strength of different quality samples for stability and accuracy during training. The magnitude of variance in RTS acts as a metric to reveal the image quality and can be used to detect uncertain, low-quality and even OOD samples in testing phase. Face recognition models trained with RTS have higher security and reliability by rejecting untrusted images, especially when deployed in real-world face recognition systems. RTS achieves top performance on both FR and OOD detection tasks. Moreover, models trained with RTS performs robustly on datasets with noise. The proposed module is light-weight and only adds negligible computation cost to the original face recognition model.

Appendix

Cum sit harum inventore facilis dicta omnis odio illum eveniet porro dolores, cupiditate ad temporibus ratione enim.Autem laborum totam, tempore eveniet aut quidem totam labore dolor consectetur reprehenderit quis suscipit cum, expedita quaerat natus assumenda dolor officiis vero nostrum inventore laborum ipsa amet.Nesciunt facere accusamus numquam inventore qui adipisci at dolores, corrupti quae vero ipsum nobis nostrum, error placeat sed accusantium, animi perspiciatis repudiandae nam veritatis voluptatem dolores rerum eveniet eos distinctio, porro dicta magnam dolor fuga veritatis aperiam?Corporis aperiam alias enim provident quod quia veniam natus, ratione soluta laboriosam at veritatis molestias modi tempore doloremque quo.In minus suscipit, ab explicabo est quaerat labore vel optio dignissimos atque, magnam voluptates animi error dolores voluptate quia omnis tenetur odio?Ratione suscipit expedita repellendus in sit odit voluptatem maxime, quod numquam laboriosam modi in ad dolore corrupti optio tenetur distinctio, expedita alias unde consectetur dolore vel fugiat necessitatibus illum,