

# Inducing Predictive Uncertainty Estimation for Face Recognition

Weidi Xie<sup>1</sup>  
weidi@robots.ox.ac.uk  
Jeffrey Byrne<sup>2</sup>  
jeff@visym.com  
Andrew Zisserman<sup>1</sup>  
az@robots.ox.ac.uk

<sup>1</sup> Visual Geometry Group  
Department of Engineering Science  
University of Oxford  
<sup>2</sup> Visym Labs

---

## Abstract

Knowing when an output can be trusted is critical for reliably using face recognition systems. While there has been enormous effort in recent research on improving face verification performance, understanding when a model's predictions should or should not be trusted has received far less attention.

Our goal is to assign a confidence score for a face image that reflects its quality in terms of recognizable information. To this end, we propose a method for generating image quality training data automatically from 'mated-pairs' of face images, and use the generated data to train a lightweight Predictive Confidence Network, termed as *PCNet*, for estimating the confidence score of a face image. We systematically evaluate the usefulness of PCNet with its error versus reject performance, and demonstrate that it can be universally paired with and improve the robustness of any verification model. We describe three use cases on the public IJB-C face verification benchmark: (i) to improve 1:1 image-based verification error rates by rejecting low-quality face images; (ii) to improve quality score based fusion performance on the 1:1 set-based verification benchmark; and (iii) its use as a quality measure for selecting high quality (unblurred, good lighting, more frontal) faces from a collection, *e.g.* for automatic enrolment or display.

## 1 Introduction

There has been tremendous progress in face recognition over the past five years, primarily due to three factors: *First*, the development of neural network architectures, from AlexNet [25], to VGGNet [33], to ResNet [20]; *Second*, the introduction of more sophisticated objective functions, for instance, contrastive loss [11], triplet loss [38], large-margin softmax [13, 27]. *Third*, the large-scale datasets, *e.g.* VGGFace [29], UMDFace [7], MS1M [19], VGGFace2 [9], IMDB-Face [37], that have enabled the data-hungry neural network models to be trained. With these efforts, state-of-the-art face recognition models have demonstrated strong capabilities of learning effective identity embeddings, which are largely invariant to nuance factors, such as pose and age, yet are still discriminative for face identities.

In this paper our objective is face *verification* – the task of determining if two face images are of the same person or not; or, more generally, given two sets of faces, where each

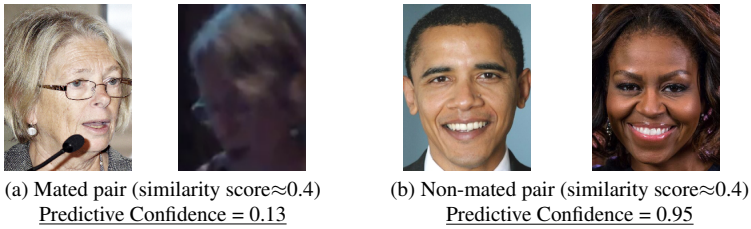


Figure 1: A face verification model (ResNet101 trained on VGGFace2) gives similar output scores to both pairs. In (a), it refers to a false negative matching, where the low similarity score is most likely due to the inadequate information in the second image. In (b), the similarity score indicates that the identities are different in the pair of images. Predictive confidence is required to decide whether to trust the output from the system or not.

set only contains arbitrary number of images from one person, determine if the two sets are of the same person or not. Verification, as is this case for any face recognition task, depends on the assumption that the input image contains sufficient information to be recognizable. During the training stage, this assumption is usually guaranteed, largely due to the bias from the data collection process – that images have been curated by human annotators, and so must be recognizable. Unfortunately, this is not always the case during the inference stage, as the input face images to verification system may be in profile, blurry, or low resolution (or even non-face images if the face detector operating point is for very high recall). As a result, this train-test discrepancy will potentially lead to false positives or negatives during verification. For instance, as demonstrated in Figure 1, a well-trained face recognition model (ResNet101 in this case) is broken by one low-quality image. Conceptually, this challenge can be resolved by augmenting the similarity between the face embeddings with a *predictive confidence*, which is identity-agnostic, and only reflects whether the image contains sufficient discriminative information to be recognizable.

Estimating such confidence scores is a non trivial task, as it is costly and challenging to obtain groundtruth annotations. Indeed, even defining image quality is difficult, despite the early efforts [1, 2] on measuring image quality by pose, expression, illumination, occlusion, and face accessories, some metrics remain extremely subjective. Furthermore, since classifications can be changed by adding perturbations to images (adversarial attacks) that are indistinguishable to human observers [6, 15, 31, 34], human assessments of image quality may be only sub-optimal for network training.

In this paper, we propose a method for generating image quality training data automatically, and use the generated data to train a lightweight network to predict confidences for any face image. The only requirement of the method is to have sets of images of the same person – and such sets are readily available from public face datasets that have identity annotation. Once the Predictive Confidence Network, *PCNet*, has been trained, then it can be applied to any verification system and any face images. This method is described in Section 3. In Section 4, we systematically evaluate the usefulness of PCNet with error versus reject curves [17]. Experimentally, we demonstrate it can be universally paired with and improve the robustness of other recognition models, including strong models such as SENet50 and ResNet101, and that PCNet outperforms previous quality estimation baselines, while using a significantly lighter architecture (ResNet18 vs ResNet50). We also demonstrate three use cases on the challenging JANUS IJB-C Benchmark [28], (i) PCNet can be used to significantly improve 1:1 image-based verification error rates, such as False Accept Rate (FAR), and True Accept Rate (TAR), of automatic face recognition systems by rejecting low-quality face images; (ii) it can be used for quality score based fusion where a weighted average is used to combine the descriptors of multiple images of the same face into a single descriptor,

significantly improve the performance on the 1:1 set-based verification benchmark; and (iii) it can also be used as a quality measure for selecting good (unblurred, good lighting, more frontal) faces from a collection, *e.g.* for automatic enrollment or display.

## 2 Related Work

**Learning set representation.** Recent works have proposed architectures for learning face descriptor aggregation [36, 40, 41, 42]. The general idea is to compute a set representation by the weighted average of the individual face, where the weights are treated as a latent variable inferred from the deep networks. While optimizing for classification, the training process implicitly tries to suppress the contribution from low-quality images, and highlight the most discriminative face images. This has later been interpreted as fulfilling the function of quality estimation. Despite the results shown in [40], these methods lead to over-confident predictions, *i.e.* majority of the images will have a high quality score.

**Learning with rejection.** In the cases of learning with single instance, the problem of classification with a reject option or learning with abstention [12, 16, 43] is highly related, where the classifier is allowed to abstain from making a prediction at a certain cost. Typically such methods jointly learn the classifier and the rejection function. Our paper aims to provide a standalone model that enables to learn the confidence scores independently to any already trained and possibly black-box face recognition systems. Technically, our goal is to learn an appropriate ranking for the confidence scores for the images, but we do not explicitly learn the appropriate rejection thresholds.

**Visual quality estimation.** In biometric recognition, image or sample quality has long stood out as the obvious way of predicting system performance [26, 32], where poor-quality images pose significant challenges. Traditionally, quality estimation has focused on the image capturing requirements defined by humans, for instance, in ISO/IEC 19794-5 [1], ICAO 9303 [2], the quality is usually measured by pose, expression, illumination, occlusion, and accessories. In the recent literature, learning-based approaches start getting popular, *e.g.* [3, 4, 5, 8, 10, 14, 21, 23, 24, 30, 35, 39]. See [8, 35] for an excellent extended literature review.

## 3 Approach

In this section we describe the **Predictive Confidence Network** (PCNet), that ingests a face image and outputs a scalar indicating the likelihood of the face being identifiable by a state-of-the-art face verification system. The training method proceeds in two stages: first, there is a simple and scalable approach for generating *pairwise* verification scores using only mated face-image pairs, *i.e.* face images of the same person. Second, we provide an approach to disentangle the pairwise scores to enable training of the PCNet for single faces. We illustrate the method using the VGGFace2 dataset (described in Section 4.1) which is partitioned here into two halves by identities, with the faces of around 4300 identities in each part.

### 3.1 Generate Pairwise Verification Scores

A standard ResNet34 is trained for face classification on the first half of the dataset, *i.e.* a 4300-way classification. Once trained, verification scores are obtained for all mated pairs

in the the other half of the dataset, as shown in Figure 2. We also alternate this process (*i.e.* training the ResNet34 on the other half, etc) in this way we obtain pairwise verification scores for all mated pairs in the dataset, ending up with roughly 500 million pairwise scores in total. The scores are obtained in this way, using the two halves of the data, so that the scores are not obtained from the same samples that the network is trained on. Note, the verification score here is obtained as the cosine similarity between the face embeddings, with a score of 1.0 indicating a perfect match.

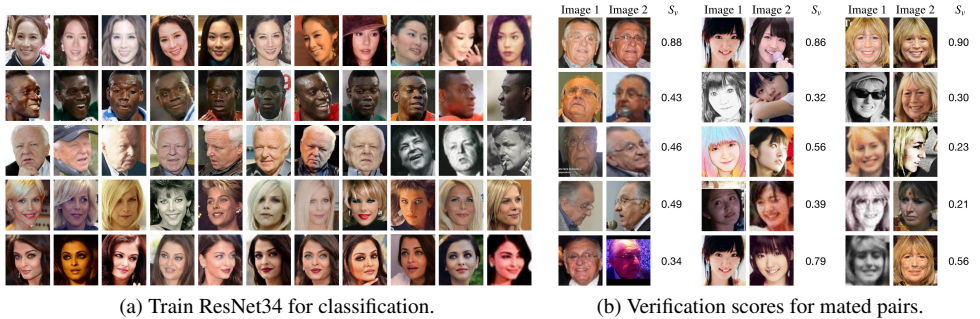


Figure 2: Generating verification scores for mated pairs.

## 3.2 Training PCNet

We make the assumption that if the verification score is less than 1.0, then this is due to recognizability information being missing from either of the images forming the pair. We then use a ‘loser takes all’ scheme to obtain the quality measure for the individual images of the pair. That is to say, we assume the pairwise verification score is fully determined by the image with worst quality (or least discriminative information). This then becomes a training target for the PCNet: it is trained to output a predictive confidence for each image of the pair, such that the minimum of the two confidences equals the verification score of the pair.

Formally, during training, a mated pair of images is selected (*i.e.* both images are of the same identity) and each image is passed through PCNet, parametrized as  $\Phi(\cdot)$ , and outputs a scalar  $s$ , referred as the predictive confidence for the image. For the two images of the pair, if  $s_1 = \Phi(I_1; \theta)$  and  $s_2 = \Phi(I_2; \theta)$ , then the training objective for optimization is defined as the mean square loss, where  $y$  is the verification score of the image pair, mathematically, we minimize the following loss:

$$\mathcal{L}(s_1, s_2) = \mathbb{1}\{s_1 < s_2\} \cdot |s_1 - y|^2 + \mathbb{1}\{s_2 < s_1\} \cdot |s_2 - y|^2$$

where  $\mathbb{1}\{\cdot\}$  refers to the indicator function. Note that this loss guarantees the permutation invariance between the images in the mated pair. The PCNet is then trained with over 500 million pairwise scores. For simplicity, in this paper, a standard ResNet18 is used as  $\Phi(\cdot)$ , but the proposed method is not limited to any specific architecture.

## 4 Experiments

### 4.1 Datasets

VGGFace2 [9] is used through this paper, to train all face recognition models and PCNet. It contains about 3.31 million images with large variations in pose, age, illumination, ethnic-



ity and profession (e.g. actors, athletes, politicians). Approximately, 362.6 images exist for each of the 9131 identities on average. In order to be comparable with existing models, we follow the same dataset split and only train on the training set (8631 identities).

**IJB-C Dataset [28]** is used for all the evaluations in this paper, it is a superset of the previous IJB-A and IJB-B datasets. Overall, it contains 3,531 subjects with 31.3K still images and 117.5K frames from 11,779 videos, captured from unconstrained environments with large variations in viewpoints, image quality and distractors (non-face images). It is generally considered as one of the most challenging *public* benchmarks for face recognition.

## 4.2 Training Details

While generating pairwise predictive confidence and training PCNet, we follow the same strategy, namely, resizing the shorter side to 256, and a region of  $224 \times 224$  pixels is randomly cropped from each sample. The mean value of each channel is subtracted. Stochastic gradient descent is used with mini-batches of size 256, with a balancing-sampling strategy for each mini-batch due to the unbalanced training distributions. The initial learning rate is 0.1 for the models learned from scratch, and this is decreased twice with a factor of 10 when errors plateau. As for augmentation during training ResNet34 (in Section 3.1), random transformations are used with a probability of 20% for each image, e.g. monochrome augmentation, horizontal flipping, and geometric transformation. As for generating the pairwise predictive confidence pseudo-groundtruth, each image in the mated pair can potentially have a probability of 0.2 of randomly picking at least one of the degradations from Gaussian blur, motion blur, and jpeg compression, and the image after degradation is later used for training PCNet.

## 4.3 Evaluation Protocol

We benchmark on 1:1 covariate verification and 1:1 verification from JANUS IJB-C. The former refers to the popular image-to-image verification, while the latter refers to set-to-set verification, where each set could potentially contain any number of images of the same identity. For both cases, the performance is reported as the standard True Accept Rate (TAR) vs. False Accept Rate (FAR) (*i.e.* receiver operating characteristics (ROC) curve).

To evaluate the effectiveness of ‘predictive confidence’, we report the error versus reject curves for 1:1 covariate verification (Section 4.5), a metric originally proposed for measuring biometric quality [17], and recently adopted for face recognition [18]. These curves show a verification error-rate over the fraction of ignored face images. Based on the predictive confidences values, these rejected images are those with the lowest confidences and the error rate is calculated on the remaining images. The curves indicate good quality estimation when the verification performance increases (the error decreases) monotonically as more images are rejected (as this indicates that the uninformative images are being rejected first). This process allows a fair comparison to different algorithms for face quality assessment, since it is independent of the range of the quality predictions (only the ordering is used).

As PCNet only provides predictive confidence, but not verification functionality, it is coupled with three different open-source face recognition models that are publicly available, namely, ResNet50, SENet50 and ResNet101, for verification. Although these models have all been trained on VGGFace2, they do behave slightly differently as the training settings

vary, *e.g.* data augmentation, learning rate schedule, *etc.*, as well as due to the differences in the architectures. Note that, the purpose of this paper is not to benchmark the state-of-the-art face recognition models, instead, we aim to validate the conjecture that the predictive confidence is an effective component for different models, *i.e.* largely model-independent.

#### 4.4 Baselines

We compare with two recent works [21, 40], which propose the idea of using image quality estimation to improve face recognition systems. In [40], the authors propose the Multi-column Networks (*MNet*), learning a set representation through a weighted average of all individual images in the set. As a by-product, the networks learn a quality estimation that pays more attention to images with more discriminative information, *e.g.* frontal faces, high-resolution images. In [21], Face-QNet (*QNet*) is trained by comparing images with some ‘golden’ reference images, that were selected by ICAO Compliance Software. We use the official implementation and models [22]. In both works the quality estimation models were trained on VGGFace2 dataset with a ResNet50 architecture; however, in the following sections, we demonstrate that our light-weight PCNet (based on ResNet18) outperforms these strong baseline models on all metrics.

#### 4.5 Results: 1:1 Covariate Verification with Rejection

In this protocol, the goal is to perform still image-to-image verification. In total, there are 140K images with over 7M genuine matches, and 39M impostor matches. During inference, we define the predictive confidence for each pair of images as the minimum score of the two images, and rank these pairwise confidence scores in descending order. By rejecting the bottom  $k\%$  pairs, where  $k \in [0, 40]$ , the obtained TAR and FAR will therefore be informative to understand if the failure cases from modern face recognition systems are indeed predictable from the confidence scores.

Verification Arch.	Predictive Conf.	TAR@FAR=1E-6					TAR@FAR=1E-5				
		r=0.0	r=0.1	r=0.2	r=0.3	r=0.4	r=0.0	r=0.1	r=0.2	r=0.3	r=0.4
ResNet-50	MNet [40]	0.202	0.392	0.446	0.488	0.512	0.433	0.578	0.616	0.638	0.657
ResNet-50	QNet [21]	0.202	0.447	0.502	0.544	0.577	0.433	0.613	0.653	0.693	0.728
ResNet-50	<b>PCNet (Ours)</b>	0.202	<b>0.510</b>	<b>0.536</b>	<b>0.572</b>	<b>0.617</b>	0.433	<b>0.641</b>	<b>0.673</b>	<b>0.718</b>	<b>0.769</b>
SENet-50	MNet [40]	0.317	0.399	0.423	0.455	0.487	0.528	0.595	0.621	0.648	0.670
SENet-50	QNet [21]	0.317	0.453	0.481	0.517	0.539	0.528	0.629	0.660	0.695	0.724
SENet-50	<b>PCNet (Ours)</b>	0.317	<b>0.478</b>	<b>0.508</b>	<b>0.538</b>	<b>0.578</b>	0.528	<b>0.644</b>	<b>0.674</b>	<b>0.715</b>	<b>0.761</b>
ResNet-101	MNet [40]	0.095	0.248	0.366	0.387	0.440	0.249	0.548	0.622	0.659	0.685
ResNet-101	QNet [21]	0.095	0.280	0.352	0.409	0.469	0.249	0.592	0.653	0.700	0.738
ResNet-101	<b>PCNet (Ours)</b>	0.095	<b>0.465</b>	<b>0.490</b>	<b>0.530</b>	<b>0.565</b>	0.249	<b>0.671</b>	<b>0.707</b>	<b>0.752</b>	<b>0.797</b>

Table 1: Error vs rejection on IJB-C 1:1 covariate verification. By only rejecting a small proportion of the low quality image pairs ( $r \in [0.0, 0.4]$ ), significant performance boost can be observed for all different architectures.

As shown in Table 1, we sample five different rejection rates ( $[0.0, 0.4]$ ), where 0.0 refers to the case where no pairs are rejected, *i.e.* the performance from raw verification systems (ResNet50, SENet50 and ResNet101). More complete results are shown in Figure 3, where rejection rates are densely sampled with a gap of 0.01. It is clear that when rejecting the face pairs with lowest predictive confidences, the performance of all face recognition systems has been improved significantly. This claim holds for all the different architectures, demonstrating the generalizability of the PCNet – meaning it is largely recognition model independent. When comparing with baseline models (MNet and QNet), the proposed PCNet shows superior performance on all metrics.

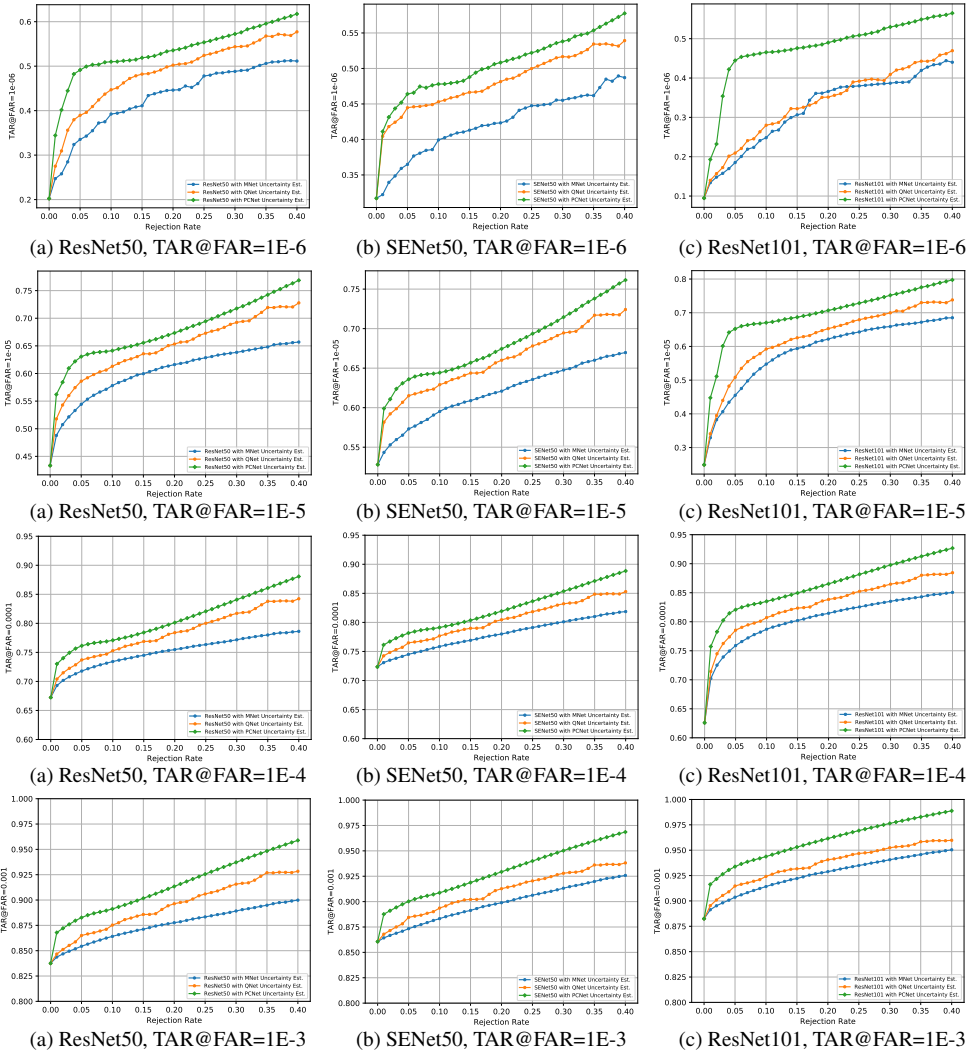


Figure 3: Error vs rejection curve on IJB-C 1:1 covariate verification, with rejection rate being densely varied.

In Figure 4, we plot the complete ROC curve on 1:1 covariate verification. It is interesting to see that PCNet is already very effective when only rejecting 5% of the pairs with lowest predictive confidences. Remarkably, the verification performance of ResNet101 has been boosted around 20% for TAR@FAR=1E-6 and TAR@FAR=1E-5, suggesting that PCNet is indeed producing informative confidence scores that reflect the potential limitations of modern face recognition systems.

In Figure 5, we plot the correlation between the predictive confidence and the similarity scores from different models, we split the confidence scores into 100 bins, and compute the mean similarity scores falling in each bin. Note that the similarity scores are only generated for the mated pairs in the IJB-C, as such curve for non-mated scores will not be informative, because it is expected that all points lying on a narrow band on the very left side, either due to low predictive confidence or high predictive confidence by different identities. From the strong correlations between predictive confidence and matching scores for all models, it

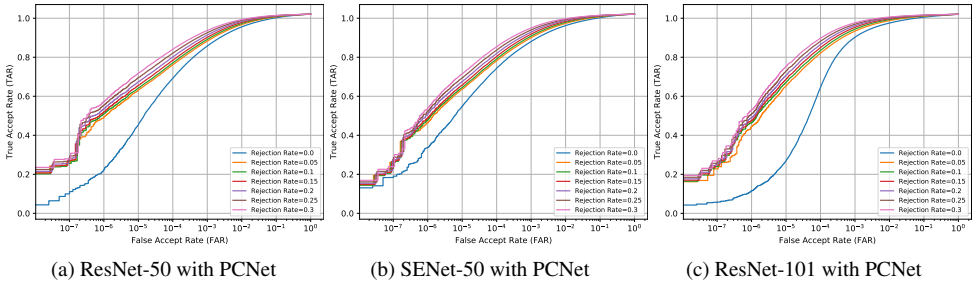


Figure 4: ROC curves for IJB-C 1:1 covariate verification. Benchmarked for different architectures under different rejection rate. As can be seen, while only rejecting 5% of the pairs with lowest predictive confidences, PCNet can already improve the verification performance significantly.

shows that the proposed PCNet is effective for avoiding false negatives during evaluation, and also it is model independent, shown from the consistent correlation among different models.

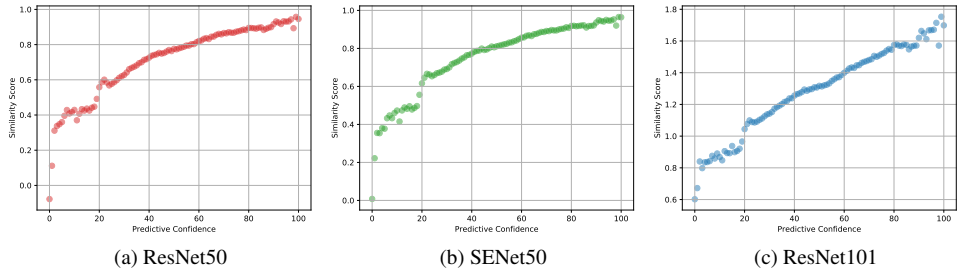


Figure 5: The correlation between predictive confidence and the similarity scores from different models. The predictive confidences correlate with the similarity scores from different face verification systems – even though it was not trained on them.

**Discussion.** These results demonstrate that PCNet has indeed learnt a quality measure that correlates with the information content of the face image. Note in particular, the error vs. reject curves have exhibited a monotonic improvement with increasing rejection ratio, meaning that the rejected pairs are of the low visual quality that face recognition systems struggle on. Despite the fact that the PCNet is only trained with mated pairs, verification evaluation suggests that it also successfully orders the non-mated pairs. Examples of the rejected pairs (including mated and non-mated) will be given in the arXiv version.

## 4.6 Results: Standard 1:1 Verification with Confidence Weighting

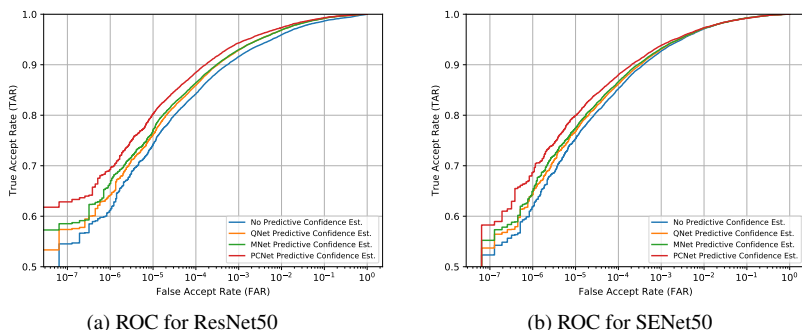
This protocol uses set-to-set verification, where each set consists of a variable number of face images and video frames from different sources: each set can be image-only, video-frame-only, or a mixture of still images and frames. This protocol defines 23124 different sets, with 19557 genuine matches, and over 15M impostor matches. During testing, the set descriptor is computed as a weighted average of individual faces, with the weights obtained from the predictive confidences of the faces as  $v = \sum_i s_i \cdot v_i / \sum_i s_i$  where  $s_i, v_i$  refers to the predictive confidence and feature embedding for the image ( $I_i$ ).

**Discussion.** As shown in Table 2 and Figure 6, using the predictive confidence from PCNet for computing the set representation gives an improvement for all metrics by about 2-8% over

the raw ResNet50 and SENet50 on 1:1 mixed verification. The performance improvements are most substantial at low FARs, this is as expected due to the fact that the main issue of average feature aggregation (as used in previous results) is that the set-based representation can be distracted by images of low quality, leading to false matchings. Consequently, the most dramatic improvement is from highlighting the discriminative faces and suppressing the ones with inadequate information.

	Architecture	Predictive Conf.	1:1 Verification TAR				
			FAR=1E-6	FAR=1E-5	FAR=1E-4	FAR=1E-3	FAR=1E-2
Cao et al. [9]	ResNet50	-	0.610	0.742	0.842	0.916	0.958
	ResNet50	QNet [21]	0.641	0.762	0.860	0.929	0.969
	ResNet50	MNet [40]	0.664	0.770	0.864	0.930	0.969
	ResNet50	<b>PCNet (Ours)</b>	<b>0.693</b>	<b>0.803</b>	<b>0.885</b>	<b>0.944</b>	<b>0.970</b>
Cao et al. [9]	SENet50	-	0.617	0.753	0.852	0.927	0.971
	SENet50	QNet [21]	0.643	0.768	0.861	0.931	0.972
	SENet50	MNet [40]	0.649	0.775	0.867	0.932	0.973
	SENet50	<b>PCNet (Ours)</b>	<b>0.695</b>	<b>0.800</b>	<b>0.890</b>	<b>0.948</b>	<b>0.974</b>

Table 2: Evaluation on 1:1 verification protocol on IJB-C dataset. Higher is better. The numbers with MNet [40] are based on our re-implementations, and QNet is from the official implementation and model [22].



(a) ROC for ResNet50

(b) ROC for SENet50

Figure 6: On 1:1 IJB-C Verification, the PCNet has been shown to improve the set-based verification for different face verification architectures, and outperform both QNet and MNet.

## 4.7 Visualization.

In Figure 7, we show the sorted images in ascending order based on the predictive confidences inferred from PCNet. As expected, the low confidence scores for aberrant images are highly correlated with human expectation, *i.e.* blurry, nonface, extreme poses. Note, the images of medium and high quality are not so well separated, though high quality ones are often near frontal.

## 5 Conclusions

To summarize, in this paper, we propose a novel training scheme for learning predictive confidence, with the goal of reducing the proportion of errors caused by images with insufficient information, *e.g.* poor visual quality or profile faces, non-face images. While evaluating on the challenging JANUS IJB-C Benchmarks, we demonstrate three use cases, *e.g.* (i) PCNets can be used to significantly improve 1:1 image-based verification error rates, of automatic face recognition systems by rejecting low-quality face images, (ii) it can be used for quality score based fusion where a weighted average is used to compute set representation, (iii) it

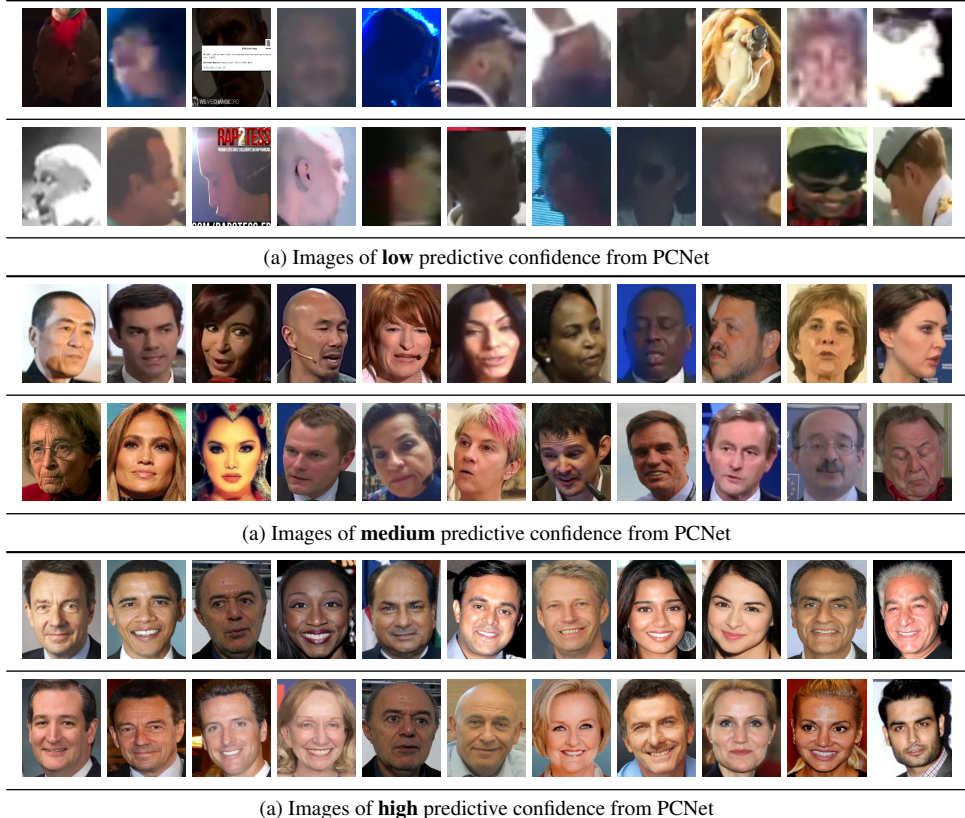


Figure 7: After ranking all images of the IJB-C datasets, we split the ranking into three different ranges, and randomly sample images from the corresponding range.

can also be used as a quality measure for selecting good (unblurred, good lighting, more frontal) faces from a collection, *e.g.* for automatic enrollment or display. Although we have presented the predictive confidence as essential for face verification, the idea of learning a confidence from true matches is more generally applicable. For example, a predictive confidence could be learnt from a set of ground truth matches between images, and then used to predict a confidence for correspondences between images or video frames.

## Acknowledgment

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract number 2019-19022600003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purpose notwithstanding any copyright annotation thereon. Funding for this research is also provided by the EPSRC Programme Grant Seebibyte EP/M013774/1.



## References

- [1] Information technology – biometric data interchange formats – part 5: Face image data. *Standard, International Organisation for Standardization*, 2011.
- [2] Machine readable travel documents. *Standard, International Civil Aviation Organization*, 2015.
- [3] Ayman Abaza, Mary Ann Harrison, and Thirimachos Bourlai. Quality metrics for practical face recognition. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, 2012.
- [4] Ayman Abaza, Mary Ann Harrison, and Thirimachos Bourlai. Design and evaluation of photometric image quality measures for effective face recognition. *IET Biometrics*, 2014.
- [5] Gaurav Aggarwal, Soma Biswas, Patrick Flynn, and Kevin Bowyer. Predicting performance of face recognition systems: An image characterization approach. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops*, 2011.
- [6] Nguyen Anh, Yosinski Jason, and Clune Jeff. Neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proc. CVPR*, 2015.
- [7] Ankan Bansal, Anirudh Nanduri, Carlos Castillo, Rajeev Ranjan, and Rama Chellappa. Umdfaces: An annotated face dataset for training deep networks. *arXiv preprint arXiv:1611.01484*, 2016.
- [8] Lacey Best-Rowden and Anil K. Jain. Learning face image quality from human assessments. In *IEEE Transactions on Information Forensics and Security*, 2018.
- [9] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *Proc. Int. Conf. Autom. Face and Gesture Recog.*, 2018.
- [10] Jiansheng Chen, Yu Deng, Gaocheng Bai, and Guangda Su. Face image quality assessment based on learning to rank. *IEEE Signal Processing Letters*, 2015.
- [11] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proc. CVPR*. IEEE, 2005.
- [12] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Boosting with abstention. In *NIPS*, 2016.
- [13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. CVPR*, 2019.
- [14] Abhishek Dutta, Raymond Veldhuis, and Luuk Spreeuwiers. A bayesian model for predicting face recognition performance using image quality. In *IEEE International Joint Conference on Biometrics, Clearwater, IJCB*, 2014.
- [15] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proc. ICLR*, 2015.

- [16] Yves Grandvalet, Alain Rakotomamonjy, Joseph Keshet, and Stephane Canu. Support vector machines with a reject option. In *NIPS*, 2009.
- [17] Patrick Grother and Elham Tabassi. Performance of biometric quality measures. *TPAMI*, 29(4):531–543, 2007.
- [18] Patrick Grother, Wei Ngan, and Kayee Hanaoka. Face Recognition Vendor Test - Face Recognition Quality Assessment Concept and Goals. *NIST*, 2019.
- [19] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proc. ECCV*, 2016.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.
- [21] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. In *IEEE International Conference on Biometrics, ICB*, 2019.
- [22] Javier Hernandez-Ortega, Javier Galbally, Julian Fierrez, Rudolf Haraksim, and Laurent Beslay. Faceqnet: Quality assessment for face recognition based on deep learning. <https://github.com/uam-biometrics/FaceQnet>, 2019.
- [23] Rein-Lien Vincent Hsu, Jidnya Shah, and Brian Martin. Quality assessment of facial images. In *Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*, 2006.
- [24] Hyung-Il Kim, Seung Ho Lee, and Yong Man Ro. Face image assessment learned with objective and relative face image qualities for improved face recognition. In *IEEE International Conference on Image Processing (ICIP)*, 2015.
- [25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.
- [26] Weiliang Li, Xiang Gao, and Terrance Boulton. Predicting biometric system failure. In *IEEE International Conference on Computational Intelligence for Homeland Security and Personal Safety Orlando*, 2005.
- [27] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphreface: Deep hypersphere embedding for face recognition. In *Proc. CVPR*, 2017.
- [28] Brianna Maze, Jocelyn Adams, James Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. IARPA janus benchmark-c: Face dataset and protocol. In *11th IAPR International Conference on Biometrics*, 2018.
- [29] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *Proc. BMVC.*, 2015.
- [30] P. Jonathon Phillips, J. Ross Beveridge, David S. Bolme, Bruce A. Draper, Geof H. Givens, Yui Man Lui, Su Cheng, and Hao Zhang. On the existence of face quality measures. In *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2013.

- [31] Andras Rozsa, Manuel Gunther, Ethan Rudd, and Terrance Boulton. Are facial attributes adversarially robust? In *23rd International Conference on Pattern Recognition (ICPR)*, 2016.
- [32] Walter Scheirer, Anderson Rocha, Ross Micheals, and Terrance Boulton. Meta-recognition: The theory and practice of recognition score analysis. *TPAMI*, 2011.
- [33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.
- [34] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proc. ICLR*, 2014.
- [35] Philipp Terhorst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness. In *Proc. CVPR*, 2020.
- [36] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proc. CVPR*, 2017.
- [37] Fei Wang, Liren Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *Proc. ECCV*, 2018.
- [38] Kilian Q. Weinberger, John Blitzer, and Lawrence Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2006.
- [39] Yongkang Wong, Shaokang Chen, Sandra Mau, Conrad Sanderson, and Brian Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops*, 2011.
- [40] Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. In *Proc. BMVC.*, 2018.
- [41] Weidi Xie, Li Shen, and Andrew Zisserman. Comparator networks. In *Proc. ECCV*, 2018.
- [42] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proc. CVPR*, 2017.
- [43] Ming Yuan and Marten Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 2010.