

On the Capacity of Face Representation

Sixue Gong, *Student Member, IEEE*, Vishnu Naresh Boddeti, *Member, IEEE*
and Anil K. Jain, *Life Fellow, IEEE*

Abstract—Face recognition is a widely used technology with numerous large-scale applications, such as surveillance, social media and law enforcement. There has been tremendous progress in face recognition accuracy over the past few decades, much of which can be attributed to deep learning based approaches during the last five years. Indeed, automated face recognition systems are now believed to surpass human performance in some scenarios. Despite this progress, a crucial question still remains unanswered: *given a face representation, how many identities can it resolve?* In other words, *what is the capacity of the face representation?* A scientific basis for estimating the capacity of a given face representation will not only benefit the evaluation and comparison of different face representation methods, but will also establish an upper bound on the scalability of an automatic face recognition system. We cast the face capacity estimation problem under the information theoretic framework of capacity of a Gaussian noise channel. By explicitly accounting for two sources of representational noise: *epistemic* (model) uncertainty and *aleatoric* (data) variability, our approach is able to estimate the capacity of any given face representation. To demonstrate the efficacy of our approach, we estimate the capacity of a 128-dimensional state-of-the-art deep neural network based face representation, FaceNet [1], and that of the classical Eigenfaces [2] representation of the same dimensionality. Our numerical experiments indicate that, (a) our capacity estimation model yields a capacity upper bound of 1×10^{12} for FaceNet and 1×10^0 for Eigenface representation at a false acceptance rate (FAR) of 5%, (b) the capacity of the face representation reduces drastically as you lower the desired FAR (for FaceNet representation, the capacity at FAR of 0.1% and 0.001% is 2×10^7 and 6×10^3 , respectively), and (c) the performance of the FaceNet representation is significantly below the theoretical limit.

Index Terms—Face Recognition, Face Representation, Channel Capacity, Gaussian Noise Channel, Bayesian Inference

1 INTRODUCTION

Face recognition has witnessed rapid progress and wide applicability in a variety of practical applications: social media, surveillance systems and law enforcement. Fueled by copious amounts of data, ever growing computational resources and algorithmic developments, current state-of-the-art face recognition systems are believed to surpass human capability in certain scenarios [3]. Despite this tremendous progress, a crucial question still remains unaddressed, *what is the capacity of a given face representation?* The face capacity here¹ is defined as the maximal number of identities that can be completely resolved² by a given face representation. Tackling this question is the central aim of this paper.

The ability to determine the capacity of a face representation affords many benefits: (a) Face representations are typically compared by their recognition performance on benchmark datasets. However, this metric of comparison is highly dependent on the complexity³ and scale of the dataset and does not showcase the full potential and limitations of the representation. Capacity reflects the discriminative power of the feature representation, consequently capacity offers an alternate data agnostic metric for comparing different representations; (b) As the deployment scale of face recognition systems grows larger (e.g., FBI face

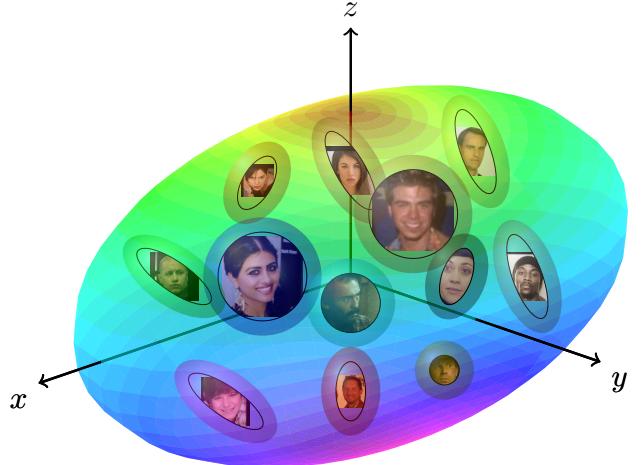


Fig. 1: An illustration of a face representation embedding, where all the faces typically lie inside the population hyper-ellipsoid. The embedding of images belonging to each identity or a class are ideally clustered into their own class-specific hyper-ellipsoids. The capacity of a face representation is the number of identities (class-specific hyper-ellipsoids) that can be packed into the population hyper-ellipsoid of the embedding within an error tolerance.

The authors are with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48864 USA e-mail: (gongsixu@msu.edu).

1. This is different from the notion of capacity of a space of functions as measured by its VC dimension.

2. Within an error tolerance.

3. Variations in facial appearance due to pose, illumination, expressions, occlusions etc.

database [4] and Aadhar [5]), it is critical to obtain reliable statistical estimates of the upper bound on the number of identities the face representation can resolve. This would allow for informed deployment of face recognition systems based on the expected scale of operation.

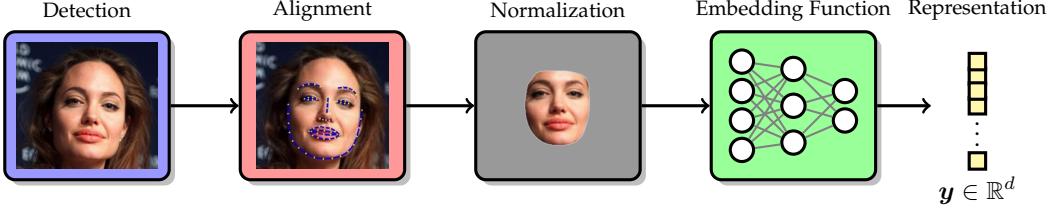


Fig. 2: A typical **face representation pipeline** comprises of face detection, alignment, normalization and representation or feature extraction. While each of these components affect the capacity of the representation, in this paper, we focus on the capacity of the embedding function that maps a high-dimensional normalized face image to a d -dimensional vector representation.

Our solution to estimate the capacity of a face representation relies on the notion of capacity that has been well studied by the information theory community in the context of wireless communication [6]. The setting, commonly referred to as the Gaussian channel, consists of a source signal $x \in \mathbb{R}^d$ that is additively corrupted by Gaussian noise $z \in \mathbb{R}^d$ to generate observations $y \in \mathbb{R}^d$. The capacity of this Gaussian channel is defined as the number of distinct source signals in the signal representation. Figure 1 provides a pictorial illustration of the geometrical structure of a face representation under the setting of a Gaussian channel. The capacity of this Gaussian channel can be computed under various assumptions for the distributions of the source signal x and the noise z .

Despite the rich theoretical understanding of the capacity of a Gaussian channel, there has been limited practical application of this theory in the context of estimating the capacity of learned embeddings like face representations, natural language processing, etc. Many challenges contribute to this limited practical utility.

- 1) Estimating the distributions of the source P_x and the noise P_z for high-dimensional embedding, such as in a face representation, is an open problem.
- 2) Accounting for the different sources of noise and reliably inferring the probability distributions, especially in high dimensions (typically, several hundred), is a very challenging task. In the context of face representations, all the components of a typical face representation pipeline (see Fig. 2), including face detection, landmarking, normalization and the mapping from a high-dimensional image to the low-dimensional embedding are potential sources of noise.
- 3) The Gaussian channel [6] setting makes independence⁴ i.e., $x \perp z$ and additivity i.e., $x + z$ assumptions. These assumptions do not hold in many practical scenarios.
- 4) Existing capacity estimates for the Gaussian channel are predicated upon assumptions on the Gaussianity of the distributions of x and z , assumptions that are again difficult to justify in many practical scenarios.

In this paper, we propose a framework that addresses the aforementioned challenges to obtain reliable estimates of the

4. $P_{x,z} = P_x P_z \quad \forall x \in \mathbb{R}^d \text{ and } \forall z \in \mathbb{R}^d$

capacity of any given face representation. We leverage recent advances in deep neural networks (DNNs) to estimate high-dimensional distributions of P_x and P_z given observed face representation y . Given an embedding function that maps normalized high-dimensional facial images to a low-dimensional vector, we model two sources of uncertainty that contribute to the noise in the embeddings: (i) uncertainty in the data, and (ii) uncertainty in the embedding function. We note that, in this paper, we only focus on the noise inherent to the embedding function and leave the noise introduced by face detection, landmarking, and normalization components for a future study. Finally, for tractability purposes, the data and noise models for our embedding are designed to conform to the Gaussian channel setting, in terms of independence, additivity and Gaussianity, and relying on the ability of deep neural networks to approximate complex non-linear functions. These, rather simplifying, assumptions enable the direct applicability of existing capacity estimates of Gaussian channels to the problem at hand. As these assumptions are relaxed in future studies, our capacity estimates will serve as a baseline. We perform empirical analysis on the Gaussianity of the embedding in the supplementary material. The key technical contributions of this paper are:

- 1) A general purpose approach for estimating high-dimensional distributions of Gaussian channel based embeddings i.e., P_y , P_x and P_z , given samples \mathbf{Y} from P_y .
- 2) A noise model P_z for facial embeddings that explicitly accounts for uncertainty due to data and the uncertainty in the parameters of the representation function.
- 3) The first practical attempt at estimating the capacity of face representations under the Gaussian channel framework.
- 4) Establishing a relationship between the capacity of a Gaussian channel and the discriminant function of a nearest neighbor classifier. Consequently, we can estimate capacity as a function of the desired operating point, in terms of the maximum desired probability of false acceptance error, of face recognition systems.
- 5) An estimate of the capacity of a state-of-the-art DNN based face representation, namely FaceNet, consisting of 128 features.

Numerical experiments suggest that our proposed

TABLE 1: A summary of face recognition systems reported in the literature.

Authors	Probe		Gallery		Training	Datasets		Performance
	# Images	# classes	# Images	# classes		Verification	Verification	
Verification (TAR @ 0.1% FAR)								
Taigman et al. [9]	4,000,000	4,000	13,113	5,749	WebD ¹	LFW		97.35%
Liao et al. [10]	8,707	4,249	1,000	1,000	N/A	LFW		41.66%
Kemelmacher-Shlizerman et al. [11]	100,000	500	690,572	1,027,060	[1]	MegaFace+FaceScrub		98.00%
Yan et al. [12]	16,028	466	116,028	466	FRGC v2.0	FRGC v2.0(Exp. 1)+SelfD ²		97.03%
Best-Rowden et al. [13]	9,074	596	4,249	4,249	COTS ³	LFW		89.00%
Wang et al. [14]	4,350	1,000	1,000	1,000	CASIA	LFW		97.52%
Schroff et al. [1]	N/A	N/A	N/A	N/A	Google	LFW		99.63%
Wen et al. [15]	N/A	N/A	N/A	N/A	WebD ¹	LFW		99.28%
Parikh et al. [16]	N/A	N/A	N/A	N/A	SelfD ²	LFW		98.95%
Retrieval (Mean Average Precision)								
Wu et al. [17]	220	N/A	1M+	N/A	N/A	WebD +LFW		44.00%
Chen et al. [18]	120	12	13,113	5,749	N/A	LFW		18.60%
Identification (Detection and Identification Rate @ RANK 1)								
Yi et al. [19]	1,196	1,196	201,196	N/A	FERET	FERET(fb)+SelfD ²		99.58%
Klare et al. [20]	50,000	50,000	50,000	50,000	PCSO	PCSO		85.00%

¹ WebD are datasets downloaded from the Internet and used to augment the gallery; different retrieval systems use their own web datasets

² SelfD are datasets collected by authors of the papers to augment the gallery

³ COTS is a commercial off the shelf face recognition system

model can provide reasonable estimates of face representation capacity, serving as a proxy for the discriminative capability of a given face representation. Applying our approach to a state-of-the-art DNN based face representation, FaceNet, yields an upper bound on the capacity of 10^8 at a false accept rate (FAR) of 0.1% i.e., the representation should have a true accept rate (TAR) of 100% at FAR of 0.1% over 10^8 subjects. However, empirically the FaceNet representation only achieves a TAR of 50% at a FAR of 0.1% on the IJB-B [7] dataset over 1,845 subjects and a TAR of 95% at a FAR of 0.1% on the LFW [8] over 5,749 subjects.

2 RELATED WORK

The subject of face recognition is as old as the field of computer vision [21]. Not surprisingly, face recognition has received tremendous attention in the computer vision and biometrics communities over the past several decades. While an exhaustive survey of the vast literature on facial feature extraction and matching is beyond the scope of this paper, we present a few notable approaches. Eigenfaces [2] and Fisherfaces [22] are among the earliest learning based approaches for face recognition that relied on second-order statistics of the data, through principal component analysis (PCA) and linear discriminant analysis (LDA), respectively. Later on, feature representations to encode the local texture in a face image became very popular for face recognition, including local binary patterns (LBP) [23] and histogram of oriented gradients (HoG) [24]. Developments in deep neural network based representation learning have contributed to massive strides in face recognition capabilities. The defining characteristic of such methods is the use of convolutional neural network (CNN) based feature extractor, a learnable embedding function comprised of several sequential linear and non-linear operators [25]. Taigman et al. [9] presented the DeepFace system, a deep CNN trained to classify faces using a dataset of 4 million examples spanning 4,000 unique identities, that demonstrated remarkable performance on the Labeled Faces in the Wild (LFW) dataset [8]. Researchers from Google [1] used a massive dataset of about 200 million

images of 8 million identities to train a CNN directly for face verification. They optimize a loss function based on triplets of images comprising a pair of similar and a pair of dissimilar faces. This model is currently the state-of-the-art for face verification and achieves the best performance on LFW and YouTube Faces dataset [26]. For the benefit of the readers, we present an overview of recent facial representation approaches in Table 1. While the focus of the majority of the work in the literature has been on the accuracy of facial matching on benchmark datasets, our goal in this paper is to characterize the maximal discriminative capacity of a given face representation at a specified error tolerance.

Information theory has been widely used in the domains of data compression and channel coding [6]. Beyond the extensive studies on the capacity estimates of communication channels [6], information theory has also been used in signal processing to derive information-theoretic limits on subspace classification. Erdogmus et al. [27] theoretically studied how information transfer through a classifier affects its performance. They illustrated a theoretical use of Renyi's definition of information, extending Fano's result, to derive an upper bound on the probability of classification error. Motivated by applications in high-dimensional signal processing, Nokleby et al. [28] derived fundamental limits on the performance of compressive linear classifiers. They identified a duality between classification (through Gaussian mixture models) and communications over non-coherent multiple-antenna channels. Lastly, in their follow-up work Nokleby et al. [29], derived tighter upper bounds on the classification of linear and affine subspaces from noisy linear features, where the subspaces are modeled by high-dimensional Gaussian distributions with approximately low-rank covariances. In contrast to the linear subspace based classifiers that are the target of these prior studies, in this paper, we seek to estimate the capacity of any given face representation embedding, including but not limited to deep neural network based face representations.

One of the key technical contributions of our paper is the reliable estimation of the underlying probability distri-

butions of our noise model P_y , P_x and P_z from samples \mathbf{y} . Gaussian Processes [30] are a popular and powerful tool in statistics that allows us to model distributions over functions, offering nice properties such as uncertainty estimates over function values, robustness to over-fitting, and principled ways for hyper-parameter tuning. A number of approaches have been proposed for modeling uncertainties in deep neural networks [31], [32], [33], [34]. Along similar lines, Kendall et al. [35] study the benefits of explicitly modeling *epistemic*⁵ (model) and *aleatoric*⁶ (data) uncertainties [36] in Bayesian deep neural networks for semantic segmentation and depth estimation tasks. Drawing inspiration from this work, we account for these two sources of uncertainties in the process of mapping a normalized facial image into a low-dimensional face representation.

Apart from the face, there are a number of other physical human traits that serve as biometric signatures for human identification, most notably fingerprints and iris. Unlike the face, capacity estimates to determine the uniqueness of these two biometric modalities have already been established. Pankanti et al. [37] derived an expression for estimating the probability of a false correspondence between minutiae-based representations from two arbitrary fingerprints belonging to two different fingers. Zhu et al. [38] later developed a more realistic model of fingerprint individuality through a finite mixture model to represent the distribution of minutiae in fingerprint images, including minutiae clustering tendencies and dependencies in different regions of the fingerprint image domain. Daugman [39] proposed an information theoretic approach to compute the capacity of IrisCode. He first developed a generative model of IrisCode based on Hidden Markov Models and then estimated the capacity of IrisCode by calculating the entropy of this generative model. To the best of our knowledge, no such capacity estimation models have been proposed in the literature for face representations. Moreover, the distinct nature of representations for fingerprint⁷, iris⁸ and face⁹ based human recognition does not allow capacity estimation approaches to carry over from one biometric modality to another. Therefore, we believe that a customized model is necessary to establish the capacity of face representations.

3 CAPACITY OF FACE REPRESENTATIONS

We first describe the setting of our problem and provide a high-level outline of our proposed approach. Our goal is to estimate the capacity of a given face representation model, dubbed *teacher*¹⁰, that is assumed to be available in the form of a black-box embedding function that maps an image s to a low-dimensional embedding \mathbf{y} . We first cast the *teacher's* face representation embedding process in the framework of a Gaussian channel (Section 3.1). We then learn a *student* model that mimics the *teacher* while also providing uncertainty estimates of the embedding, enabling us to estimate

- 5. Uncertainty due to lack of information about a process.
- 6. Uncertainty stemming from the inherent randomness of a process.
- 7. An unordered collection of minutiae points.
- 8. A binary representation, called the iris code.
- 9. A fixed-length vector of real values.
- 10. We adopt the terminology of teacher-student models from the model compression community [40].

the parameterized probability distributions, P_y , P_x and P_z , inherent to the Gaussian channel (Section 3.2). Then we empirically estimate the capacity of the *student* model by leveraging the theoretical framework for the capacity of a Gaussian noise channel (Section 3.3). Finally, we establish a relationship between the Gaussian channel and the discriminant function of a nearest neighbor classifier, allowing us to obtain capacity estimates as a function of the probability of false acceptance (Section 3.4). A pictorial outline of our proposed face representation capacity estimation approach is shown in Fig. 3.

3.1 Face Representation Model

A face representation model M is a parametric embedding function that maps a face image s of identity c to a, typically, lower dimensional vector space $\mathbf{y} \in \mathbb{R}^d$, i.e., $\mathbf{y} = f_M(s; \theta)$, where θ is the set of parameters of the embedding function. For example, in the case of a linear embedding function like Principal Component Analysis (PCA), the parameter set θ would represent the eigenvectors. And, in the case of a deep neural network based non-linear embedding function, θ represents the parameters of the deep neural network.

The aforementioned face embedding process can be approximately cast within the framework of a Gaussian noise channel as follows. Face representations \mathbf{y} of an image s from the *teacher* are modeled as observations of a true underlying embedding \mathbf{x} that is corrupted by noise \mathbf{z} . The nature of the relationship between these entities is determined by the assumptions of a Gaussian channel, namely, (i) additivity of the noise i.e., $\mathbf{y} = \mathbf{x} + \mathbf{z}$, (ii) independence of the true embedding and the additive noise, i.e., $\mathbf{x} \perp \mathbf{z}$, and (iii) all entities, \mathbf{y} , \mathbf{x} and \mathbf{z} follow a Gaussian distribution, i.e., $P_x \sim \mathcal{N}(\mu_x, \Sigma_x)$, $P_z \sim \mathcal{N}(\mathbf{0}, \Sigma_z)$ and $P_y \sim \mathcal{N}(\mu_y, \Sigma_y)$. Statistical estimates of these parameterized distributions will enable us to compute the capacity of the *teacher* face representation model as described in Section 3.3.

But, before we proceed further, we would like to point out a major limitation of our face representation model, namely the Gaussianity assumption of the distribution P_x . For a given black-box face representation, in practice, the embeddings could potentially lie on an arbitrary and unknown low-dimensional manifold. Approximating this manifold through a normal distribution potentially overestimates the support of the embedding in \mathbb{R}^d , resulting in an over-estimation of the capacity of the representation. Similarly, the Gaussianity assumption of the noise distribution P_z is another limitation of our model. The potentially non-linear nature of the embedding function would, in general, result in an arbitrary noise distribution.

While this particular representation model suffers from the aforesaid drawbacks, it affords the practical advantage of trading-off the realism and complexity of the theoretical model and the complexity of the inferential algorithm¹¹. Specifically, our representation model allows us to compensate for the lack of theoretical machinery for modeling the manifolds induced by non-linear high-dimensional embedding functions by leveraging the functional approximation

¹¹. Similar simplifying assumptions were made by Pankanti et al. [37] for fingerprints and Daugman [39] for iris. Zhu et al. [38] later relaxed the model assumptions made by Pankanti et al.

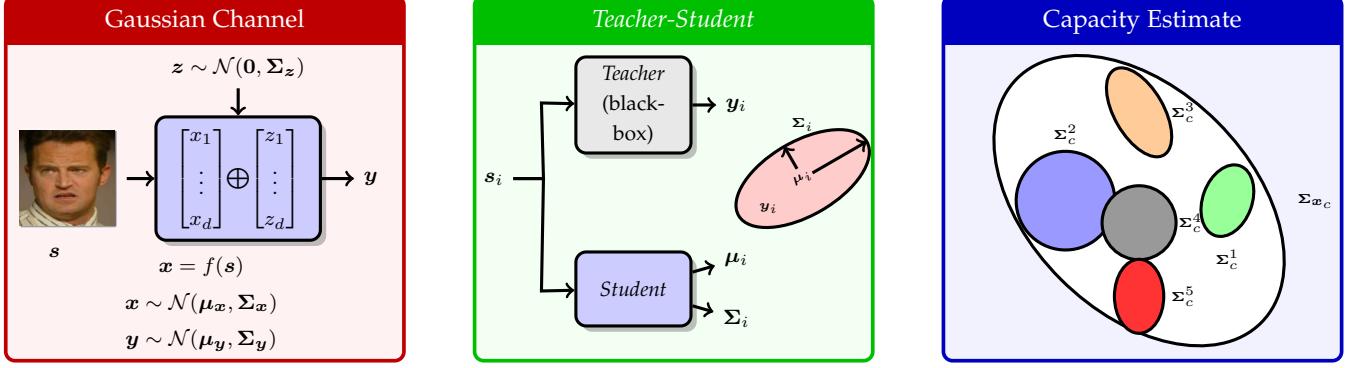


Fig. 3: Overview of Face Representation Capacity Estimation: We cast the face representation process in the information theoretic framework of a Gaussian noise channel. Given a black-box *teacher* face representation model, we learn a *student* model that mimics this *teacher* while being able to provide explicit estimates of the uncertainty (noise) in the embedding due to the data and the model. The uncertainty estimates at the image level are leveraged to generate empirical approximations of the probability distributions of the various components of the Gaussian channel at the identity level, which in turn facilitates an empirical estimate of the capacity of the *teacher* face representation.

capabilities, flexibility and computational efficiency of deep neural networks as feed-forward inference machines.

3.2 Estimating Uncertainties in Representations

We first build a probabilistic model for the space of noisy embeddings \mathbf{y} generated by a black-box facial representation model (*teacher*) M_t with parameters θ .

$$\begin{aligned} p(\mathbf{y}|\mathbf{S}^*, \mathbf{Y}^*) &= \int p(\mathbf{y}|\mathbf{x}, \mathbf{S}^*, \mathbf{Y}^*)p(\mathbf{x}|\mathbf{S}^*, \mathbf{Y}^*)d\mathbf{x} \quad (1) \\ &= \int \int p(\mathbf{y}|\mathbf{x}, \theta)p(\theta|\mathbf{S}^*, \mathbf{Y}^*)p(\mathbf{x}|\mathbf{S}^*, \mathbf{Y}^*)d\theta d\mathbf{x} \end{aligned}$$

where $\mathbf{Y}^* = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ and $\mathbf{S}^* = \{s_1, \dots, s_N\}$ are the training samples to estimate the model parameters θ , $p(\mathbf{y}|\mathbf{x}, \theta)$ is the *aleatoric* (data) uncertainty given a set of parameters, $p(\theta|\mathbf{S}^*, \mathbf{Y}^*)$ is the *epistemic* (model) uncertainty in the parameters given the training samples and $p(\mathbf{x}|\mathbf{S}^*, \mathbf{Y}^*) \sim \mathcal{N}(\mu_g, \Sigma_g)$ is the Gaussian approximation of the underlying manifold of noiseless embeddings. Furthermore, we assume that the true mapping between the image s and the noiseless embedding x is a deterministic but unknown function i.e., $\mathbf{x} = f(s)$.

The black-box nature of the *teacher* model however only provides $\mathcal{D} = \{s_i, \mathbf{y}_i\}_{i=1}^N$, pairs of facial images s_i and their corresponding noisy embeddings \mathbf{y}_i , a single sample from the distribution $p(\mathbf{y}|\mathbf{S}^*, \mathbf{Y}^*)$. Therefore, we learn a *student* model M_s with parameters w to mimic the *teacher* model. Specifically, the *student* model approximates the data dependent *aleatoric* uncertainty $p(\mathbf{y}_i|s_i, w) \sim \mathcal{N}(\mu_i, \Sigma_i)$, where $\mu_i = \mathbf{x}_i = f(s_i)$ represents the data dependent mean noiseless embedding and Σ_i represents the data dependent uncertainty around the mean. This *student* is an approximation of the unknown underlying probabilistic *teacher*, by which an input image s generates noisy embeddings \mathbf{y} of ideal noiseless embeddings \mathbf{x} , for a given set of parameters w , i.e., $p(\mathbf{y}_i|s_i, w) \approx p(\mathbf{y}_i|\mathbf{x}_i, \theta)$. Finally, we employ a variational distribution to approximate the *epistemic* uncertainty of the *teacher* i.e., $p(w|\mathbf{S}^*, \mathbf{Y}^*) \approx p(\theta|\mathbf{S}^*, \mathbf{Y}^*)$.

Learning: Given pairs of facial images and their corresponding embeddings from the *teacher* model, we learn a *student*

model to mimic the outputs of the *teacher* for the same inputs in accordance to the probabilistic model described above. We use parameterized functions, $\mu_i = f(s_i; w_\mu)$ and $\Sigma_i = f(s_i; w_\Sigma)$ to characterize the *aleatoric* uncertainty $p(\mathbf{y}_i|s_i, w)$, where $w = \{w_\mu, w_\Sigma\}$. We choose deep neural networks, specifically convolutional neural networks as our functions $f(\cdot; w_\mu)$ and $f(\cdot; w_\Sigma)$. For the *epistemic* uncertainty, while many deep learning based variational inference [31], [41], [42] approaches have been proposed, we use the simple interpretation of dropout as our variational approximation [31]. Practically, this interpretation simply characterizes the uncertainty in the deep neural network weights w through a Bernoulli sampling of the weights.

We learn the parameters of our probabilistic model $\phi = \{w_\mu, w_\Sigma, \mu_g, \Sigma_g\}$ through maximum-likelihood estimation i.e., minimizing the negative log-likelihood of the observations $\mathbf{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$. This translates to optimizing a combination of loss functions:

$$\min_{\phi} \quad \mathcal{L}_s + \lambda \mathcal{L}_g + \gamma \mathcal{L}_{r_s} + \delta \mathcal{L}_{r_g} \quad (2)$$

where λ , γ and δ are the weights for the different loss functions, $\mathcal{L}_{r_s} = \frac{1}{2N} \sum_{i=1}^N \|\Sigma_i\|_F^2$ and $\mathcal{L}_{r_g} = \frac{1}{2} \|\Sigma_g\|_F^2$ are the regularization terms and \mathcal{L}_s is the loss function of the *student* that captures the log-likelihood of a given noisy representation \mathbf{y}_i under the distribution $\mathcal{N}(\mu_i, \Sigma_i)$.

$$\mathcal{L}_s = \frac{1}{2} \sum_{i=1}^N \ln |\Sigma_i| + \frac{1}{2} \text{Trace} \left(\sum_{i=1}^N \Sigma_i^{-1} [(\mathbf{y}_i - \mu_i)(\mathbf{y}_i - \mu_i)^T] \right) \quad (3)$$

\mathcal{L}_g is the log-likelihood of the manifold of noiseless embeddings under the approximation by a multi-variate normal distribution $\mathcal{N}(\mu_g, \Sigma_g)$.

$$\mathcal{L}_g = \frac{N}{2} \ln |\Sigma_g| + \frac{1}{2} \text{Trace} \left(\Sigma_g^{-1} \sum_{i=1}^N [(\mathbf{y}_i - \mu_g)(\mathbf{y}_i - \mu_g)^T] \right) \quad (4)$$

For computational tractability we make a simplifying assumption on the covariance matrix Σ by parameterizing it as a diagonal matrix i.e., the off-diagonal elements are set to zero. This parametrization corresponds to independence assumptions on the uncertainty along each dimension of the

embedding. The sparse parametrization of the covariance matrix yields two computational benefits in the learning process. Firstly, it is sufficient for the *student* to predict only the diagonal elements of the covariance matrix. Secondly, positive semi-definitiveness constraints on a diagonal matrix can be enforced simply by forcing all the diagonal elements of the matrix to be non-negative. To enforce non-negativity on each of the diagonal variance values, we predict the log variance, $l_j = \log \sigma_j^2$. This allows us to re-parameterize the *student* likelihood (Eq. 3) as a function of l :

$$\mathcal{L}_s = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^d l_i^j + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^d \frac{(y_i^j - \mu_i^j)^2}{\exp(l_i^j)} \quad (5)$$

Similarly, we reparameterize the likelihood (Eq.4) of the noiseless embedding as a function of l_g , the log variance along each dimension.

$$\mathcal{L}_g = \frac{1}{2} \sum_{j=1}^d l_g^j + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^d \frac{(y_i^j - \mu_g^j)^2}{\exp(l_g^j)} \quad (6)$$

The regularization terms are also reparameterized as, $\mathcal{L}_{r_s} = \frac{1}{2N} \sum_{i=1}^N \sum_{j=1}^d \exp(l_i^j)$ and $\mathcal{L}_{r_g} = \frac{1}{2} \sum_{j=1}^d \exp(l_g^j)$.

Since the noise \mathbf{z} is a normal distribution with zero-mean in our Gaussian channel model, $E(\mathbf{y}) = E(\mathbf{x})$. Therefore, we empirically estimate μ_g as $\mu_g = \frac{1}{N} \sum_{i=1}^N \mathbf{y}_i$. We estimate the other parameters $\phi = \{\mathbf{w}_\mu, \mathbf{w}_\Sigma, \Sigma_g\}$ through stochastic gradient descent [43]. The gradients of the parameters are computed by backpropagating [44] the gradients of the outputs through the network.

Inference: The *student* model that has been learned can now be used to infer the uncertainty in the embeddings of the original *teacher* model. For a given facial image s , the *aleatoric* uncertainty can be predicted by a feed-forward pass of the image s through the network i.e., $\mu = f(s, \mathbf{w}_\mu)$ and $\Sigma = f(s, \mathbf{w}_\Sigma)$. The *epistemic* uncertainty can be approximately estimated through Monte-Carlo integration over different samples of model parameters w . In practice the parameter sampling is performed through the use of dropout at inference. In summary, the total uncertainty in the embedding of each facial image s is estimated by performing Monte-Carlo integration over a total of T evaluations as follows:

$$\hat{\mu}_i = \frac{1}{T} \sum_{t=1}^T \mu_i^t \quad (7)$$

$$\hat{\Sigma}_i = \frac{1}{T} \sum_{t=1}^T (\mu_i^t - \hat{\mu}_i) (\mu_i^t - \hat{\mu}_i)^T + \frac{1}{T} \sum_{t=1}^T \Sigma_i^t \quad (8)$$

where μ_i^t and Σ_i^t are the predicted *aleatoric* uncertainty for each feed-forward evaluation of the network.

3.3 Capacity From Uncertainty Estimates

The goal of this paper is to estimate the capacity of a given *teacher* face representation, where capacity is defined as the number of unique identities that can be resolved by the representation. Given \mathbf{x}_c and \mathbf{y}_c , the respective noiseless population embedding and the noisy embeddings of a

representative class, the capacity of the face representation channel is defined as,

$$C = \max_{p(\mathbf{x}_c)} I(\mathbf{x}_c; \mathbf{y}_c), \quad (9)$$

where $I(\mathbf{x}; \mathbf{y}) = \int \int p(\mathbf{x}, \mathbf{y}) \log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}$ is the mutual information between the random variables \mathbf{x} and \mathbf{y} . For general distributions, it may not be possible to analytically calculate the capacity of a channel. However, for the Gaussian channel that we consider in this paper, analytical expressions for the capacity do exist. In fact, the capacity of this channel has been extensively studied (we refer the reader to [6] for a more in-depth treatment of this topic) across various relaxations of the Gaussian distributions, namely, isotropic Gaussian, axis aligned Gaussian and finally the more general case of anisotropic Gaussian distribution with colored noise, the scenario that is most relevant to our model. Under this general case, the capacity of the face representation Gaussian channel is given by:

$$\begin{aligned} C &= \max_{p(\mathbf{x}_c)} I(\mathbf{x}_c; \mathbf{y}_c) \\ &= \max_{\text{Trace}(\Sigma_{\mathbf{x}_c}) \leq P} \frac{1}{2} \log \frac{|\Sigma_{\mathbf{x}_c} + \Sigma_{\mathbf{z}_c}|}{|\Sigma_{\mathbf{z}_c}|} \\ &= \max_{\text{Trace}(\Sigma_{\mathbf{x}_c}) \leq P} \frac{1}{2} \log \frac{|U_{\mathbf{z}_c}^T \Sigma_{\mathbf{x}_c} U_{\mathbf{z}_c} + \Lambda_{\mathbf{z}_c}|}{|\Lambda_{\mathbf{z}_c}|} \\ &= \max_{\text{Trace}(\Sigma_{\bar{\mathbf{x}}_c}) \leq P} \frac{1}{2} \log \frac{|\Sigma_{\bar{\mathbf{x}}_c} + \Lambda_{\mathbf{z}_c}|}{|\Lambda_{\mathbf{z}_c}|} \end{aligned} \quad (10)$$

where $\Sigma_{\mathbf{x}_c}$ is the covariance matrix of the noiseless population embedding of classes, $\Sigma_{\mathbf{z}_c}$ is the covariance matrix of the embedding noise of a representative class, $U_{\mathbf{z}_c}$ and $\Lambda_{\mathbf{z}_c}$ are the matrices with the eigenvectors and the eigenvalues of $\Sigma_{\mathbf{z}_c}$ respectively. The capacity of this channel is maximized when $\Sigma_{\bar{\mathbf{x}}_c}$ is diagonalized. Under the assumption of axis aligned Gaussian distribution (i.e., off-diagonal elements of covariance matrix are zero) the expression for the capacity of this channel simplifies as,

$$C = \max_{\sum_{i=1}^d \sigma_{\mathbf{x}_c}^i \leq P} \frac{1}{2} \sum_{i=1}^d \log \left(1 + \frac{\sigma_{\mathbf{x}_c}^i}{\sigma_{\mathbf{z}_c}^i} \right) \quad (11)$$

where $\sigma_{\mathbf{x}_c}^i$ is the variance of the i -th dimension of the noiseless population embedding and $\sigma_{\mathbf{z}_c}^i$ is the variance of the i -th dimension of the embedding noise of a representative class.

While we were able to extract uncertainty estimates of each individual image in Section 3.2, we now need to estimate the distributions $P_{\mathbf{x}_c}$ and $P_{\mathbf{z}_c}$, the distributions for the noiseless population embedding of classes and the embedding noise of a representative class, respectively. We approximate each of these distributions by multi-variate normal distributions, $\mathbf{x}_c \sim \mathcal{N}(\mu_{\mathbf{x}_c}, \Sigma_{\mathbf{x}_c})$ and $\mathbf{z}_c \sim \mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{z}_c})$.

We empirically estimate the parameters of these distributions as follows. The mean of the population embedding is computed as $\mu_{\mathbf{x}_c} = \frac{1}{C} \sum_{c=1}^C \hat{\mu}^c$, where $\hat{\mu}^c = \frac{1}{N_c} \sum_{i=1}^{N_c} \hat{\mu}_i^c$. The covariance of the population embedding $\Sigma_{\mathbf{x}_c}$ is estimated as,

$$\Sigma_{\mathbf{x}_c} = \frac{1}{C} \sum_{c=1}^C \left[\hat{\Sigma}^c + (\hat{\mu}^c - \mu_{\mathbf{x}_c})(\hat{\mu}^c - \mu_{\mathbf{x}_c})^T \right] \quad (12)$$

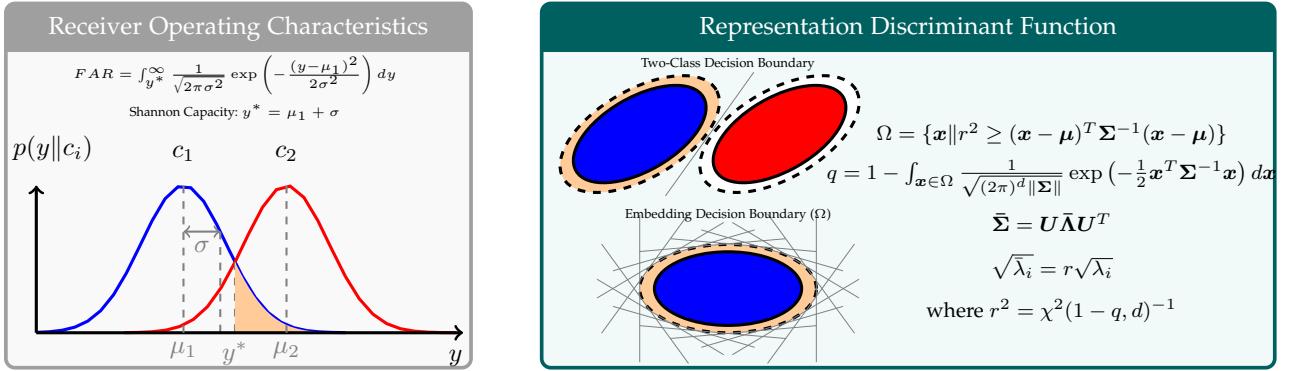


Fig. 4: Decision Theory and Capacity: We illustrate the relation between capacity and the discriminant function corresponding to a nearest neighbor classifier. **Left:** Depiction of the notion of decision boundary and probability of false accept between two identical one dimensional Gaussian distributions. Shannon's definition of capacity corresponds to the decision boundary being one standard deviation away from the mean. **Right:** Depiction of the decision boundary induced by the discriminant function of nearest neighbor classifier. Unlike in the definition of Shannon's capacity, the size of the ellipsoidal decision boundary is determined by the maximum acceptable false accept rate. The probability of false acceptance can be computed through the cumulative distribution function of a $\chi^2(r^2, d)$ distribution.

where $\hat{\boldsymbol{\Sigma}}^c = \frac{1}{N_c} \sum_{i=1}^{N_c} \hat{\boldsymbol{\Sigma}}_i^c$. Along the same lines, the covariance of the embedding noise of a given class c is estimated as,

$$\boldsymbol{\Sigma}_{\mathbf{z}_c} = \frac{1}{N_c T} \sum_{i=1}^{N_c} \sum_{t=1}^T [(\boldsymbol{\mu}_i^t - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{\mu}_i^t - \hat{\boldsymbol{\mu}}_i)^T + \boldsymbol{\Sigma}_i^t]$$

We use these estimates of $\boldsymbol{\Sigma}_{\mathbf{x}_c}$ and $\boldsymbol{\Sigma}_{\mathbf{z}_c}$ in Eq. 10 to compute the capacity of the face representation.

3.4 Decision Theory and Model Capacity

The capacity of the face representation, in terms of the Gaussian channel, described thus far can be alternatively motivated from a geometrical perspective. Recall that our face representation space is composed of two components: the noiseless population embedding of the classes approximated by a multi-variate Gaussian distribution and the embedding noise of each class approximated by a homoscedastic multi-variate Gaussian distribution. Under these settings, the decision boundaries between the classes¹² that minimizes the classification error rate are determined by discriminant functions [45]. As illustrated in Fig. 4, for a two-class problem, the discriminant function is a hyper-plane in \mathbb{R}^d with this optimal hyper-plane being equidistant from both the classes. Moreover, the separation between the classes determines the operating point and hence the probability of false acceptance. In the multi-class setting the optimal discriminant function is the surface encompassed by all the pairwise hyper-planes, which asymptotically reduces to a high-dimensional hyper-ellipsoid. The extent of this enclosing hyper-ellipsoid can be determined by the desired operating point in terms of the maximal error probability of false acceptance.

Under the multi-class setting, the capacity estimation problem is equivalent to the geometrical problem of ellipse packing, which seeks to estimate the maximum number

of small hyper-ellipsoids that can be packed into a larger hyper-ellipsoid. In the context of face representations the small hyper-ellipsoids correspond to the class-specific enclosing hyper-ellipsoids as described above while the large hyper-ellipsoid corresponds to the entire space spanned by all the classes. The volume V of a hyper-ellipsoid corresponding to a Mahalanobis distance $r^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ with covariance matrix $\boldsymbol{\Sigma}$ is given by the following expression, $V = V_d |\boldsymbol{\Sigma}|^{\frac{1}{2}} r^d$, where V_d is the volume of the d -dimensional hypersphere. An upper bound on the capacity of the face representation can be computed simply as the ratio of the volumes of the population and the class-specific hyper-ellipsoids,

$$\begin{aligned} C &\leq \left(\frac{V_{\mathbf{x}_c, \mathbf{z}_c}}{V_{\mathbf{z}_c}} \right) \\ &= \left(\frac{V_d |\boldsymbol{\Sigma}_{\mathbf{x}_c} + \boldsymbol{\Sigma}_{\mathbf{z}_c}|^{\frac{1}{2}} r_{\mathbf{x}_c}^d}{V_d |\boldsymbol{\Sigma}_{\mathbf{z}_c}|^{\frac{1}{2}} r_{\mathbf{z}_c}^d} \right) \\ &= \left(\frac{|\boldsymbol{\Sigma}_{\mathbf{x}_c} + \boldsymbol{\Sigma}_{\mathbf{z}_c}|^{\frac{1}{2}} r_{\mathbf{x}_c}^d}{|\boldsymbol{\Sigma}_{\mathbf{z}_c}|^{\frac{1}{2}} r_{\mathbf{z}_c}^d} \right) \\ &= \left(\frac{|\bar{\boldsymbol{\Sigma}}_{\mathbf{x}_c, \mathbf{z}_c}|^{\frac{1}{2}}}{|\bar{\boldsymbol{\Sigma}}_{\mathbf{z}_c}|^{\frac{1}{2}}} \right) \end{aligned} \quad (13)$$

where $V_{\mathbf{x}_c, \mathbf{z}_c}$ is the volume of population hyper-ellipsoid and $V_{\mathbf{z}_c}$ is the volume of the class-specific hyper-ellipsoid. The size of the population hyper-ellipsoid $r_{\mathbf{x}_c}$ is chosen such that a desired fraction of all the classes lie within the hyper-ellipsoid and $r_{\mathbf{z}_c}$ determines the size of the class-specific hyper-ellipsoid. $\bar{\boldsymbol{\Sigma}}_{\mathbf{x}_c, \mathbf{z}_c}$ and $\bar{\boldsymbol{\Sigma}}_{\mathbf{z}_c}$ are the effective sizes of the enclosing population and class-specific hyper-ellipsoids respectively. For each of the hyper-ellipsoids the effective radius along the i -th principal direction is $\sqrt{\lambda_i} = r \sqrt{\lambda_i}$, where $\sqrt{\lambda_i}$ is the radius of the original hyper-ellipsoid along the same principal direction.

This geometrical interpretation of the capacity reduces to the Shannon capacity in Eq. 10 when $r_{\mathbf{x}_c} = r_{\mathbf{z}_c}$. Consequently, in this instance, the choice of $r_{\mathbf{x}_c}$ for the pop-

12. In the case of face recognition, each class is an identity (subject) and the number of classes corresponds to the number of identities.

ulation hyper-ellipsoid implicitly determines the boundary of separation between the classes and hence the operating false accept rate (FAR) of the embedding. For instance, when computing the Shannon capacity of the face representation choosing $r_{\mathbf{x}_c}$ such that 95% of the classes are enclosed within the population hyper-ellipsoid would implicitly correspond to operating at a FAR of 5%.

However, very often, practical face recognition systems need to operate at different false accept rates, dictated by the desired level of security. The geometrical interpretation of the capacity described in Eq. 13 directly enables us to compute the representation capacity as a function of the desired operating point as determined by its corresponding false accept rate. The size of the population hyper-ellipsoid $r_{\mathbf{x}_c}$ will be determined by the desired fraction of classes to enclose or alternatively other geometric shapes like the minimum volume enclosing hyper-ellipsoid or the maximum volume inscribed hyper-ellipsoid of a finite set of classes, both of which correspond to a particular fraction of the population distribution. Similarly, the desired false accept rate q determines the size of the class-specific hyper-ellipsoid $r_{\mathbf{z}_c}$.

Let $\Omega = \{\mathbf{x} \mid r^2 \geq (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\}$ be the enclosing hyper-ellipsoid. Without loss of generality, assuming that the class-specific hyper-ellipsoid is centered at the origin, the false accept rate q can be computed as,

$$q = 1 - \int_{\mathbf{x} \in \Omega} \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left(-\frac{\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}}{2}\right) d\mathbf{x} \quad (14)$$

Reparameterizing the integral as $\mathbf{y} = \boldsymbol{\Sigma}^{-\frac{1}{2}} \mathbf{x}$, we have $\Omega = \{\mathbf{y} \mid r^2 \geq \mathbf{y}^T \mathbf{y}\}$ and,

$$q = 1 - \int_{\mathbf{y} \in \Omega} \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{\mathbf{y}^T \mathbf{y}}{2}\right) d\mathbf{y} \quad (15)$$

where $\{y_1, \dots, y_n\}$ are independent standard normal random variables. The Mahalanobis distance r^2 is distributed according to the $\chi^2(r^2, d)$ distribution with d degrees of freedom and $1 - q$ is the cumulative distribution function of $\chi^2(r^2, d)$. Therefore, given the desired FAR q , the corresponding Mahalanobis distance $r_{\mathbf{z}_c}$ can be obtained from the inverse CDF of the $\chi^2(r_{\mathbf{z}_c}^2, d)$ distribution. Along the same lines, the size of the population hyper-ellipsoid $r_{\mathbf{x}_c}$ can be estimated from the inverse CDF of the $\chi^2(r_{\mathbf{x}_c}^2, d)$ distribution given the desired fraction of classes to encompass.

4 NUMERICAL EXPERIMENTS

In this section we will evaluate the efficacy of the proposed capacity model. We estimate the capacity of a state-of-the-art deep neural network based face representation model. Furthermore, we estimate the capacity of the face representation using multiple datasets, corresponding to different practical scenarios.

4.1 Datasets

We first provide a brief description of the face datasets that we use, both for learning the *teacher* and *student* models as well as for estimating the capacity of the *teacher*. Figure 5 shows a few examples of the kind of faces in each dataset.

PCSO: This dataset is a subset of a larger collection of mugshot images (not in the public domain) acquired from the Pinellas County Sheriffs Office (PCSO), comprising of 1,447,607 images of 403,619 subjects. We use a subset of 84,164 images from 10,000 subjects for face verification and capacity estimation.

LFW [8]: This dataset is a collection of 13,233 face images of 5,749 subjects, downloaded from the web. The images in this dataset exhibit limited variations in pose, illumination, and expression, since only faces that could be detected by the Viola-Jones face detector [46] were included in the dataset. One limitation of this dataset is that only 1,680 subjects among the total of 5,749 subjects have more than one face image.

CASIA [47]: This dataset is a large collection of labeled images downloaded from the web (based on names of famous personalities) typically used for training deep neural networks. It consists of 494,414 images across 10,575 subjects, with an average of about 500 face images per subject. This dataset is used for training both the *teacher* and *student* models.

IJB-A [48]: IARPA Janus Benchmark-A (IJB-A) contains 500 subjects with a total of 25,813 images (5,399 still images and 20,414 video frames), an average of 51 images per subject. Compared to the LFW and CASIA datasets, the IJB-A dataset is more challenging due to: i) full pose variation making it difficult to detect all the faces using a commodity face detector, ii) a mix of images and videos, and iii) wider geographical variation of subjects. The face locations are provided with the IJB-A dataset (and used in our experiments when needed).

IJB-B [7]: IARPA Janus Benchmark-B (IJB-B) dataset is a superset of the IJB-A dataset consisting of 1,845 subjects with a total of 76,824 images (21,798 still images and 55,026 video frames from 7,011 videos), an average of 41 images per subject. Images in this dataset are labeled with ground truth bounding boxes and other covariate meta-data such as occlusions, facial hair and skin tone. A key motivation for the IJB-B dataset is to make the face database less constrained compared to the IJB-A dataset and have a more uniform geographic distribution of subjects across the globe in comparison to IJB-A.

4.2 Face Representation Model

We consider the capacity of two different face representation models: (i) the state-of-the-art deep neural network based FaceNet introduced by Schroff et al. [1], and (ii) the classical PCA based EigenFaces [2] representation of image pixels. These two representations are illustrative of the two extremes of various face representations proposed in the literature with FaceNet providing close to state-of-the-art recognition performance. The FaceNet representation is based on a non-linear multi-layered deep convolutional network architecture introduced by Szegedy et al. [49]. The EigenFaces representation, in contrast, is a linear model for representing faces. However, unlike the original FaceNet model, both of our models are trained using the entire CASIA dataset. Figure 6a provides a complete description of the inception-resnet [49] network architecture of the *teacher* model, the basis of the FaceNet representation.



Fig. 5: Example images from each of the five datasets considered in our paper in increasing order of complexity. The IJB-A and IJB-B datasets exhibit large variations in pose, expressions and illumination. The LFW and CASIA datasets, however, only have limited variations in pose, illumination and expression, but LFW consists of only very few images per identity, on average. The PCSO dataset (not in the public-domain) is a mugshot dataset of frontal face images over a large and diverse set of identities.

Since the *student* model is purposed to mimic the *teacher* model, we base the *student* network architecture on the *teacher's*¹³ architecture with a few notable exceptions. First, we introduce dropout before every convolutional layer of the network, including all the convolutional layers of the inception [50] and residual [51] modules. Second, the last layer of the network is modified to generate two outputs μ and Σ instead of the noisy embedding y , the output of the *teacher*. Figure 6b provides a pictorial representation of the *student* network architecture.

4.3 Implementation Details

Before we train the *teacher* and *student* networks, the face images are pre-processed and normalized to a canonical face image. The faces are detected and normalized using the joint face detection and alignment system introduced by Zhang et al. [52]. Given the facial landmarks, the faces are normalized to a canonical image of size 182×182 from which RGB patches of size 160×160×3 are extracted as the input to the *teacher* and *student* networks. During training these patches are cropped randomly from the normalized image and also flipped horizontally at random for data augmentation.

The *teacher* is trained through a combination of the softmax loss for classification, and the center loss [15] that minimizes the intra-class distance between the embeddings. Training is performed through stochastic gradient descent with Nesterov Momentum 0.9 and weight decay 0.0005. We use a batch size of 16, a learning rate of 0.01 dropping by a factor of 2 every 20 epochs and is trained for about 100 epochs. The *teacher* model also uses dropout (with dropout probability set to 0.8) at training, but only in the final fully-connected classification layer. As is standard practice, we do not use dropout on the *teacher* model during inference.

The *student* is trained to minimize the loss function defined in Eq. 2, where the hyper-parameters are chosen through cross-validation. Training is performed through

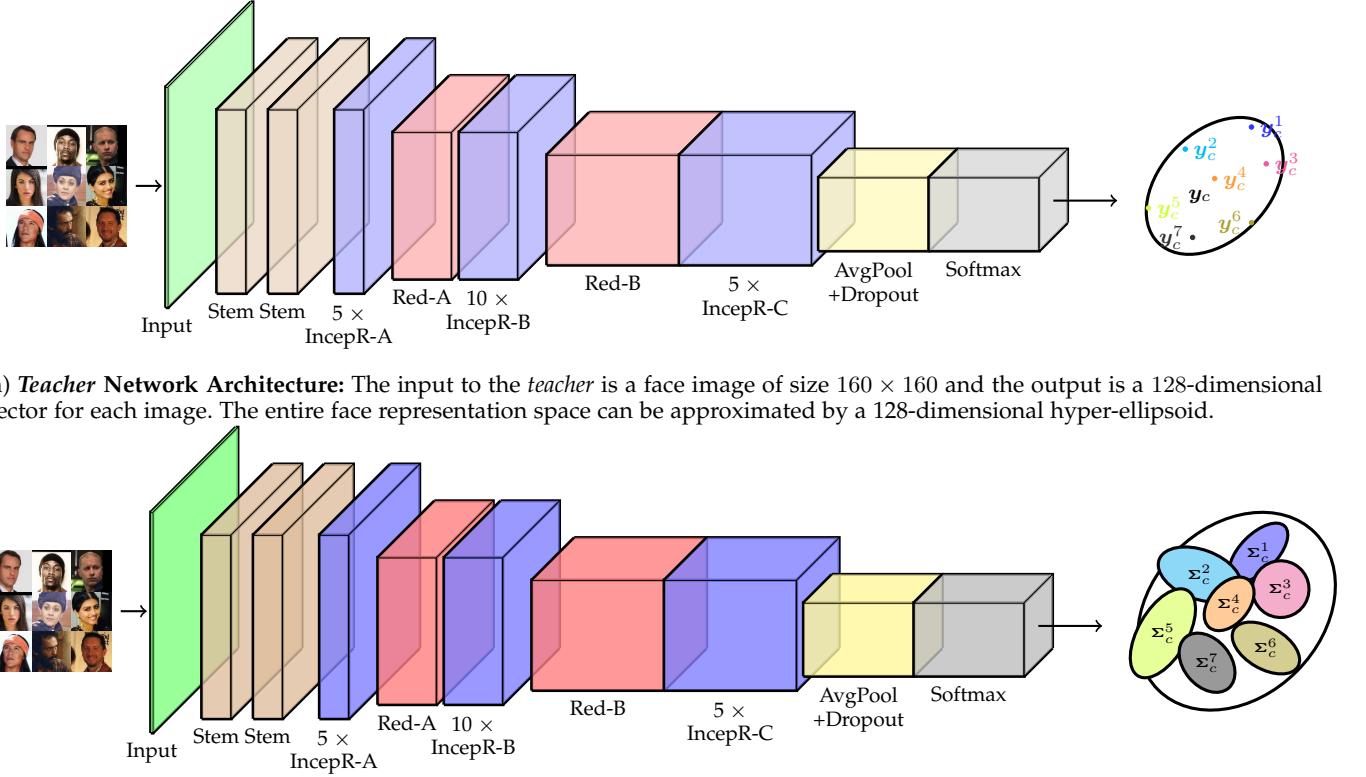
13. In the scenario where the *teacher* is a black-box model, the design of the student network architecture needs more careful consideration but it also affords more flexibility. See Fig. 3 for an illustration of this process.

stochastic gradient descent with Nesterov Momentum 0.9 and weight decay 0.0005. We use a batch size of 16, a learning rate of 0.01 that is dropped by a factor of 2 every 20 epochs. We observed that it is sufficient to train the *student* model for about 100 epochs for convergence. The *student* model includes dropout with a probability of 0.05 after each convolutional layer and with a probability of 0.8 after the final fully-connected classification layer. At inference each image is passed through the *student* network 1,000 times as a way of performing Monte-Carlo integration through the space of network parameters $\{w_\mu, w_\Sigma\}$. These sampled outputs are used to empirically estimate the mean and covariance of the image embedding.

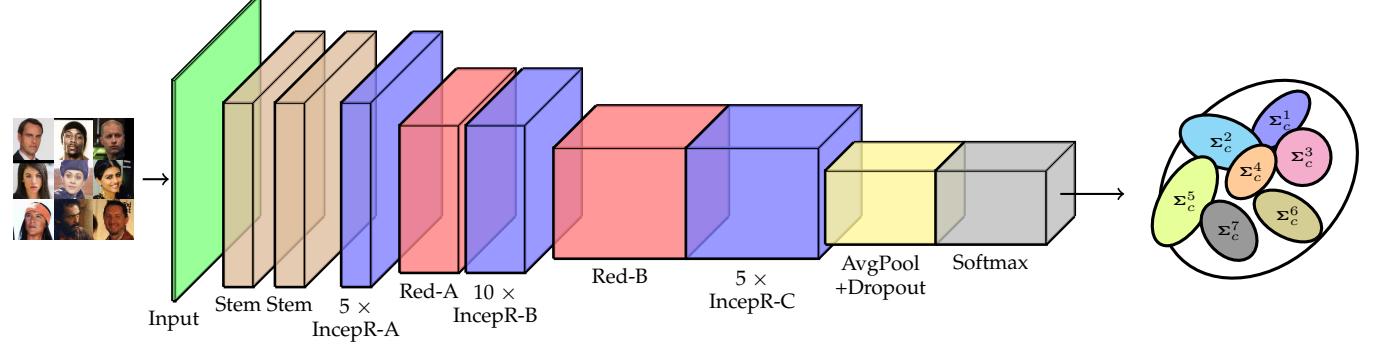
4.4 Face Recognition Experiments

We evaluate and compare the performance of the *teacher* and *student* models on the four test datasets that we consider, namely, LFW, PCSO, IJB-A and IJB-B. To evaluate the *student* model we estimate the face representation through Monte-Carlo integration. We pass each image through the *student* model 1,000 times to extract $\{\mu_i, \Sigma_i\}_{i=1}^{1000}$ and compute $\mu = \frac{1}{1000} \sum_{i=1}^{1000} \mu_i$ as the representation. Following standard practice, we match a pair of representations through the nearest neighbor classifier on the normalized feature vectors i.e., by computing the euclidean distance $d_{ij} = \|\mu_i - \mu_j\|_2^2$ between the normalized representations $\mu_i = \frac{x_i}{\|x_i\|_2}$, where x_i is the un-normalized representation from either the *teacher* or *student* model. We note that the capacity model considered in this paper does not take into account this normalization process, the projection of the the un-normalized representations in \mathbb{R}^d onto the surface of a unit hyper-sphere in \mathbb{R}^d , leading to overestimates of the face representation capacity. We leave the topic of relaxing this simplifying assumption, to obtain more accurate capacity estimates, to a future study.

We evaluate the face representation models on the LFW unconstrained face dataset, using two protocols: the standard LFW protocol and the BLUFR protocol [10]. Since there is no standard predefined protocol for evaluation on the PCSO dataset, we use the BLUFR protocol. Lastly, the protocol for the IJB-A and IJB-B datasets defines matching



(a) **Teacher Network Architecture:** The input to the *teacher* is a face image of size 160×160 and the output is a 128-dimensional vector for each image. The entire face representation space can be approximated by a 128-dimensional hyper-ellipsoid.



(b) **Student Network Architecture:** The input to the *student* is a face image of size 160×160 and the output is a 128-dimensional hyper-ellipsoid of the uncertainty of the representation for the corresponding image. This entire face representation can be approximated by a 128-dimensional population hyper-ellipsoid.

Fig. 6: Face Representation Architectures: The network architectures for estimating the representation capacity. (a) The Inception-ResNet [49] architecture of the *teacher* face representation model. This architecture is the current state-of-the-art network for face representations. It was designed to represent a 160×160 face image by a 128-dimensional feature vector. (b) We adopt the same Inception-ResNet architecture for the *student* network. The *student* network, however, differs from the *teacher* network in the following ways: (1) it has two outputs: the mean and the covariance of the uncertainty of the embedding, and (2) incorporates dropout after every convolutional layer in the network.

between templates, where each template is composed of possibly multiple images of the class. Following the protocol in [14], we define the match score between templates as the average of the match scores between all pairs of images in the two templates.

Figure 7 and Table 2 report the performance of the *teacher* and *student* models, both FaceNet and Eigenfaces, on each of these datasets at different operating points. We make the following observations: (1) The performance of DNN based representation on PCSO and LFW, consisting largely of frontal face images with minimal pose variations and facial occlusions, is comparable to the state-of-the-art. However, its performance on IJB-A and IJB-B, datasets with large pose variations, is lower than state-of-the-art approaches. This is due to the fact that unlike these methods we do not fine-tune the DNN model on the IJB-A and IJB-B training sets since our goal in the paper is estimate the capacity of a generic face representation as opposed to achieving the best verification performance on each individual datasets. (2) Our results indicate that the *student* models are able to mimic the *teacher* models very well as demonstrated by the similarity of their recognition accuracies as well as the receiving operating curves, and (3) As expected, the

DNN significantly outperforms the PCA based Eigenfaces representation in terms of recognition accuracy.

4.5 Capacity Estimates

Shannon Capacity: We estimate the capacity of the face representations by evaluating Eq. 10. For each of the datasets we empirically determine the shape and size of the population embedding Σ_{x_c} and the class-specific hyper-ellipsoids Σ_{z_c} . These quantities are computed through the predictions obtained by sampling the weights (w_μ, w_Σ) of the model, via dropout. We obtain 1,000 such predictions for a given image, by feeding the image through the network a 1,000 different times with dropout. For robustness against outliers we only consider classes with at least two images per class for LFW and five images per class for all the other datasets for the capacity estimates.

Although the expressions in Eq. 10 and Eq. 13 assume that the shape of the class-specific embedding and the population embedding are hyper-ellipsoids, we obtain capacity estimates for different modeling assumptions on the shape of these entities. For instance the shapes could also be modeled as hyper-spheres corresponding to a diagonal covariance matrix with the same variance in each dimension. We

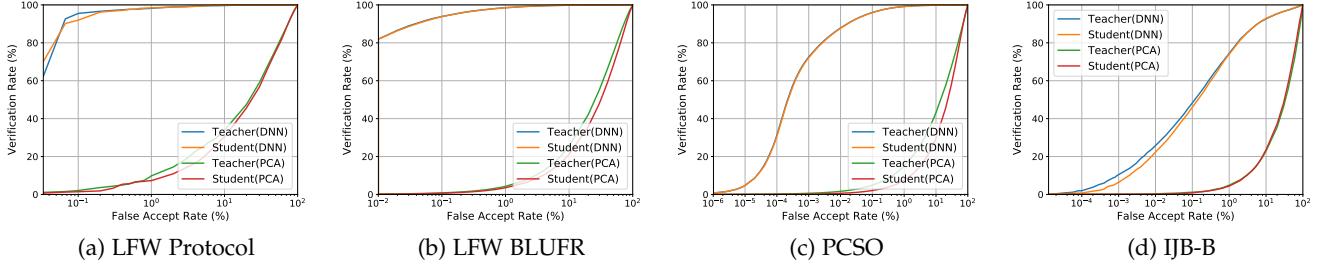


Fig. 7: Face recognition performance of the *teacher* and *student* models on the various datasets. We report the performance of both DNN and PCA based face representations, (a) face verification on the LFW dataset evaluated through the standard protocol, (b) face verification on the LFW dataset evaluated through the BLUFR protocol, (c) face verification on a subset of the PCSO dataset (84,164 images of 10,000 classes), (d) face verification on the IJB-B dataset.

TABLE 2: Face Recognition Results

Dataset	Teacher: DNN		Student: DNN		Teacher: PCA		Student: PCA		State-of-the-Art	
	0.1% FAR	1% FAR	0.1% FAR	1% FAR						
LFW	95.50	98.20	91.97	98.47	2.07	9.80	1.43	7.27	N/A	99.7
LFW (BLUFR)	93.90	98.51	90.68	97.22	0.81	4.3	0.64	3.55	98.7	N/A
PCSO	96.01	99.16	95.94	99.16	4.89	15.03	2.07	13.49	99.6	99.7
IJB-A	45.92	70.26	44.35	70.36	0.31	1.87	0.17	1.39	90.6	95.2
IJB-B	48.31	74.47	46.04	73.96	1.16	4.87	0.84	4.39	70.0	83.0

TABLE 3: Capacity of Face Representation Model

Dataset	Examples	Noise	Hypersphere				Hyper-ellipsoid (Axis Aligned)		Hyper-ellipsoid	
			Model	DNN	PCA	DNN	PCA	DNN	DNN	PCA
PCSO		Min	1.5×10^{37}	1.6×10^0	9.0×10^{34}	1.6×10^0	9.8×10^{39}	1.6×10^0	9.8×10^{39}	1.6×10^0
		Mean	3.9×10^{19}	1.0×10^0	2.1×10^4	1.0×10^0	1.6×10^{21}	1.0×10^0	1.6×10^{21}	1.0×10^0
		Median	9.5×10^{19}	1.0×10^0	9.0×10^{18}	1.0×10^0	4.4×10^{30}	1.0×10^0	4.4×10^{30}	1.0×10^0
		Max	4.0×10^0	5.9×10^{-1}	9.5×10^1	5.9×10^{-1}	1.0×10^{19}	5.9×10^{-1}	1.0×10^{19}	5.9×10^{-1}
LFW		Min	8.4×10^{36}	1.7×10^0	9.1×10^{35}	1.7×10^0	3.6×10^{39}	1.7×10^0	3.6×10^{39}	1.7×10^0
		Mean	8.7×10^{18}	1.0×10^0	7.4×10^5	9.9×10^{-1}	5.0×10^{18}	9.9×10^{-1}	5.0×10^{18}	9.9×10^{-1}
		Median	3.7×10^{19}	1.0×10^0	9.0×10^{20}	1.0×10^0	4.7×10^{31}	1.0×10^0	4.7×10^{31}	1.0×10^0
		Max	3.9×10^4	5.9×10^{-1}	1.8×10^4	5.9×10^{-1}	1.3×10^{17}	5.9×10^{-1}	1.3×10^{17}	5.9×10^{-1}
CASIA		Min	7.0×10^{27}	1.4×10^0	3.2×10^{27}	1.7×10^0	3.6×10^{34}	1.4×10^0	3.6×10^{34}	1.4×10^0
		Mean	1.7×10^{10}	1.0×10^0	2.5×10^3	1.0×10^0	5.0×10^{13}	1.0×10^0	5.0×10^{13}	1.0×10^0
		Median	2.1×10^{10}	1.0×10^0	2.2×10^{10}	1.0×10^0	6.8×10^{22}	1.0×10^0	6.8×10^{22}	1.0×10^0
		Max	1.0×10^1	6.7×10^{-1}	2.8×10^1	6.7×10^{-1}	7.5×10^{11}	6.7×10^{-1}	7.5×10^{11}	6.7×10^{-1}
IJB-A		Min	3.7×10^{21}	7.8×10^0	5.8×10^{21}	7.8×10^0	8.2×10^{29}	7.8×10^0	8.2×10^{29}	7.8×10^0
		Mean	3.4×10^9	1.0×10^0	2.2×10^4	9.9×10^{-1}	6.6×10^{13}	9.9×10^{-1}	6.6×10^{13}	9.9×10^{-1}
		Median	2.0×10^9	1.0×10^0	1.5×10^9	1.0×10^0	1.1×10^{21}	1.0×10^0	1.1×10^{21}	1.0×10^0
		Max	8.1×10^3	6.6×10^{-1}	1.5×10^3	6.6×10^{-1}	3.3×10^{12}	6.6×10^{-1}	3.3×10^{12}	6.6×10^{-1}
IJB-B		Min	1.3×10^{27}	1.5×10^0	9.1×10^{26}	1.5×10^0	3.1×10^{33}	1.5×10^0	3.1×10^{33}	1.5×10^0
		Mean	5.2×10^9	1.0×10^0	2.6×10^4	9.9×10^{-1}	2.2×10^{14}	9.9×10^{-1}	2.2×10^{14}	9.9×10^{-1}
		Median	2.7×10^9	1.0×10^0	5.7×10^9	1.0×10^0	5.0×10^{23}	1.0×10^0	5.0×10^{23}	1.0×10^0
		Max	8.0×10^2	5.8×10^{-1}	1.0×10^3	5.8×10^{-1}	8.4×10^{12}	5.8×10^{-1}	8.4×10^{12}	5.8×10^{-1}

generalize the hyper-sphere model to an axis aligned hyper-ellipsoid corresponding to a diagonal covariance matrix with possibly different variances along each dimension. For the sake of modeling efficiency we make the same modeling assumption for both the global shape of the embedding and the embedding shape of each class.

The capacity estimates of the face representations, under the Gaussian noise channel setting, are dependent on the size of a canonical class-specific hyper-ellipsoid (the capacity expression Eq. 10 depends on Σ_{z_c}), that is representative of a typical subject among all possible subjects the face representation system is going to be used for. However,

the hyper-ellipsoid corresponding to each class could potentially be of a different size. For instance, in Fig. 1 each class-specific hyper-ellipsoid is of a different size, orientation and shape. Defining or identifying a canonical subject, from among all possible identities, is in itself a challenging task and beyond the scope of this paper. We overcome this practical problem by reporting the capacity for different choices of classes, i.e., ranging from classes with very low intra-class variability and classes with very high intra-class variability. We report estimates of the capacity for different datasets and for each dataset select classes with the minimum, mean, median and maximum hyper-ellipsoid volume as our canonical



Fig. 8: Example images of classes that correspond to different sizes of the class-specific hyper-ellipsoids, based on the FaceNet representation, for each of the five datasets considered in the paper. **Top Row:** Images of the class with the largest class-specific hyper-ellipsoid for each database. Notice that in the case of database with predominantly frontal faces (PCSO and LFW), large variations in facial appearance lead to the greatest uncertainty in the class representation. On more challenging datasets (CASIA, IJB-A, IJB-B), the face representation exhibits most uncertainty due to pose variations. **Bottom Row:** Images of the class with the smallest class-specific hyper-ellipsoid for each database. As expected, across all the datasets, frontal face images with the minimal change in appearance result in the least amount of uncertainty in the class representation. More examples like these can be found in the supplementary material.

class-specific hyper-ellipsoid. Datasets whose class distribution is similar to the distribution of the data that was used to train the face representation, are expected to exhibit low intra-class uncertainty, while datasets with classes that are out of the training distribution can potentially have high intra-class uncertainty, and consequently lower capacity. Figure 8 show examples of the images corresponding to the lowest and highest intra-class variability in each dataset.

Table 3 reports the Shannon capacity of the DNN and PCA based face representation estimated on various datasets and across different modeling assumptions. We make the following observations from our numerical results, *DNN and PCA*: The capacity of the FaceNet representation is significantly higher than the capacity of Eigenfaces, with the former and later being of the order of $\approx 10^{12}$ and 10^0 respectively. The Eigenfaces representation based on linear projections of the raw pixel values is unable to scale beyond a handful of identities, while FaceNet representation based on learned non-linear functions is able to resolve significantly more number of identities. The relative difference in the capacity is also reflected in the vast difference in the verification performance between the two representations.

Capacity Estimates: The upper bound on the capacity estimate of the FaceNet model in constrained scenarios (LFW and PCSO) is $\approx 10^{18}$ and $\approx 10^{12}$ in unconstrained environments (CASIA and IJB) under the general model of a hyper-ellipsoid. Therefore, theoretically, the representation should be to resolve 10^{18} and 10^{12} subjects with a true acceptance rate (TAR) of 100% at a false acceptance rate (FAR) of 5% under the constrained and unconstrained operational settings, respectively. While this capacity estimate is significantly greater than the population of the earth, in practice, the performance of the representation is lower than the theoretical performance, about 95% across only

10,000 subjects in the constrained and only 50% across 1,845 subjects in the unconstrained scenarios. These results suggest a significant room for improvement in the empirical performance of face recognition systems, especially under unconstrained scenarios. The relative order of the capacity estimates mimics the relative order of the verification accuracy on these datasets.

Data Bias: Even within each dataset, the capacity estimates exhibit significant variations depending on the choice of the size of the class-specific hyper-ellipsoid. Whilst these variations are a consequence of the simplicity of our capacity estimation model and the inhomogeneity of distribution of samples from each class, they help establish an upper bound on the capacity of the representation. Empirically, we have observed that classes with the highest capacity (smallest hyper-ellipsoid) are typically classes with very few images and very little variation in facial appearance. Similarly, classes with high intra-class uncertainty, with low capacity estimates, are typically classes with a very large number of images spanning a wide range of variations in pose, expression, illumination conditions etc., variations that one can expect under any real-world deployments of face recognition systems. Coupled with the fact that the capacity of the face representation is estimated from a very small sample of the population (less than 11,000 subjects), we argue that capacity estimates with large intra-class uncertainty (within the datasets considered in this paper) are an accurate representation of a canonical subject's face representation in unconstrained real-world deployments of face recognition systems.

Gaussian Distribution Parameterization: The capacity estimates of the general hyper-ellipsoid are significantly higher than the capacity estimates of the reduced approximations, hyper-sphere (isotropic Gaussian) and axis-aligned hyper-

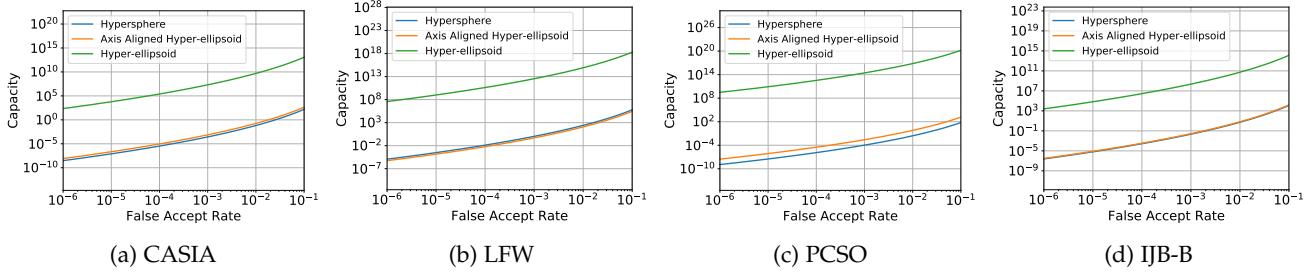


Fig. 9: For each of the datasets under consideration, we estimate the capacity of the deep neural network at different false accept rates corresponding to different operating points. Under the limit, the capacity tends to zero as the FAR tends to zero. Similarly, the capacity tends to ∞ as the FAR tends to 1.0.

ellipsoid. At the same time, the isotropic and the axis-aligned hyper-ellipsoid approximations result in very similar capacity estimates.

Capacity at Specified FAR: We extend the point estimates of Shannon’s capacity described so far to establish capacity as a function of different operating points, as defined by different false accept rates. We define r_{x_c} and r_{z_c} corresponding to the desired operating points and evaluate Eq. 13. In all our experiments we choose r_{x_c} to encompass 95% of the classes within the population hyper-ellipsoid. Different false accept rates define different decision boundary contours that, in turn, define the size of the class-specific hyper-ellipsoid. Figure 9 shows how the capacity of the representation changes as a function of the false accept rates (FAR) for different datasets. We note that at the operating point of $FAR = 0.1\%$, the capacity of the maximum face representation is $\approx 10^{13}$ in the constrained and $\approx 10^7$ in the unconstrained case. However, at stricter operating points (FAR of 0.001% or 0.0001%), that is more meaningful at larger scales of operation [11], the capacity of even the FaceNet representation is significantly lower ($\approx 10^3$ for IJB) than the typical desired scale of operation of face recognition systems.

5 CONCLUSION

In this paper, we make the first attempt to estimate the capacity of any given face representation. We approximate the face representation process as a Gaussian noise channel and explicitly account for the different sources of uncertainty in the representation. We proposed an efficient approach to infer the states of this representation channel and estimate its capacity under the information theoretic notion of capacity of a Gaussian channel. We also extended our model to establish a relation between the capacity of the Gaussian noise channel and the false accept rate of a nearest neighbor classifier. Finally, we empirically estimated the capacity, under the Gaussian noise channel, of two face representations: a state-of-the-art deep neural network, FaceNet and classical Eigenfaces. Numerical results yield an upper bound of 10^{13} for FaceNet and 10^0 for Eigenfaces at a FAR of 5% and a large gap between the theoretical and empirical verification performance of the representations. At lower FAR of 0.001% or 0.0001%, the FaceNet capacity drops-off significantly to $\approx 10^3$ under unconstrained scenarios, impairing the scalability of the face representation.

As face recognition technology makes rapid strides in performance and witnesses wider adoption, quantifying the capacity of a given face representation is an important problem, both from a theoretical as well as from a practical perspective. However, due to the challenging nature of the technical problem we make simplifying assumptions, for tractability purposes, in this first attempt. Our experimental results demonstrate that even this simplified model is able to provide reasonable capacity estimates of a state-of-the-art face representation. Relaxing the assumptions of the approach presented here is an exciting direction of future work, leading to more realistic capacity estimates.

REFERENCES

- [1] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.
 - [2] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *CVPR*, 1991.
 - [3] C. Lu and X. Tang, "Surpassing human-level face verification performance on lfw with gaussianface," in *AAAI*, 2015.
 - [4] U. S. G. A. Office, "Face recognition technology: Fbi should better ensure privacy and accuracy," 2016. [Online]. Available: <http://www.gao.gov/assets/680/677098.pdf>
 - [5] UIDAI, "Aadhar," 2017. [Online]. Available: <https://portal.uidai.gov.in/uidwebportal/dashboard.do>
 - [6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.
 - [7] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen *et al.*, "Iarpa janus benchmark-b face dataset," in *CVPRW*, 2017.
 - [8] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.
 - [9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014.
 - [10] S. Liao, Z. Lei, D. Yi, and S. Z. Li, "A benchmark study of large-scale unconstrained face recognition," in *IJCB*, 2014.
 - [11] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard, "The megaface benchmark: 1 million faces for recognition at scale," in *CVPR*, 2016.
 - [12] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Towards incremental and large scale face recognition," in *IJCB*, 2011.
 - [13] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain, "Unconstrained face recognition: Identifying a person of interest from a media collection," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2144–2157, 2014.
 - [14] D. Wang, C. Otto, and A. K. Jain, "Face search at scale: 80 million gallery," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1122 – 1136, 2017.
 - [15] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*. Springer, 2016.

- [16] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.
- [17] Z. Wu, Q. Ke, J. Sun, and H.-Y. Shum, "Scalable face image retrieval with identity-based quantization and multireference reranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1991–2001, 2011.
- [18] B.-C. Chen, Y.-Y. Chen, Y.-H. Kuo, and W. H. Hsu, "Scalable face image retrieval using attribute-enhanced sparse codewords," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1163–1173, 2013.
- [19] D. Yi, Z. Lei, Y. Hu, and S. Z. Li, "Fast matching by 2 lines of code for large scale face recognition systems," *arXiv preprint arXiv:1302.7180*, 2013.
- [20] B. F. Klare, A. Blanton, and B. Klein, "Efficient face retrieval using synecdoches," in *IJCB*, 2014.
- [21] A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," *Pattern Recognition Letters*, vol. 79, pp. 80–105, 2016.
- [22] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
- [23] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *ECCV*, 2004.
- [24] A. Albiol, D. Monzo, A. Martin, J. Sastre, and A. Albiol, "Face recognition using hog–ebgm," *Pattern Recognition Letters*, vol. 29, no. 10, pp. 1537–1543, 2008.
- [25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.
- [26] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR*, 2011.
- [27] D. Erdogmus and J. C. Principe, "Lower and upper bounds for misclassification probability based on renyi's information," *The Journal of VLSI Signal Processing*, vol. 37, no. 2, pp. 305–317, 2004.
- [28] M. Nokleby, R. Calderbank, and M. R. Rodrigues, "Information-theoretic limits on the classification of gaussian mixtures: Classification on the grassmann manifold," in *Information Theory Workshop*, 2013, pp. 1–5.
- [29] M. Nokleby, M. Rodrigues, and R. Calderbank, "Discrimination on the grassmann manifold: Fundamental limits of subspace classifiers," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 2133–2147, 2015.
- [30] C. E. Rasmussen and C. K. Williams, *Gaussian Processes for Machine Learning*. MIT Press Cambridge, 2006, vol. 1.
- [31] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *arXiv preprint arXiv:1506.02142*, vol. 2, 2015.
- [32] ——, "A theoretically grounded application of dropout in recurrent neural networks," in *NIPS*, 2016.
- [33] ——, "Bayesian convolutional neural networks with bernoulli approximate variational inference," *arXiv preprint arXiv:1506.02158*, 2015.
- [34] Y. Gal, "Uncertainty in deep learning," Ph.D. dissertation, PhD Thesis, University of Cambridge, 2016.
- [35] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *arXiv preprint arXiv:1703.04977*, 2017.
- [36] A. Der Kiureghian and O. Ditlevsen, "Aleatory or epistemic? does it matter?" *Structural Safety*, vol. 31, no. 2, pp. 105–112, 2009.
- [37] S. Pankanti, S. Prabhakar, and A. K. Jain, "On the individuality of fingerprints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1010–1025, 2002.
- [38] Y. Zhu, S. C. Dass, and A. K. Jain, "Statistical models for assessing the individuality of fingerprints," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 391–401, 2007.
- [39] J. Daugman, "Information theory and the iriscode," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 2, pp. 400–409, 2016.
- [40] J. Ba and R. Caruana, "Do deep nets really need to be deep?" in *NIPS*, 2014.
- [41] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *NIPS*, 2015.
- [42] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," *arXiv preprint arXiv:1505.05424*, 2015.
- [43] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.
- [44] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Cognitive Modeling*, vol. 5, no. 3, p. 1, 1988.
- [45] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. John Wiley & Sons, 2012.
- [46] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
- [47] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.
- [48] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *CVPR*, 2015.
- [49] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.
- [50] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.
- [51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
- [52] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.



Sixue Gong received her bachelor degree from Communication University of China in 2015. Since 2016, she has been working towards a Ph.D. degree in the Department of Computer Science and Engineering at Michigan State University. Her research interests include pattern recognition, image processing, and computer vision, with applications to face recognition.



Vishnu Naresh Boddetti is currently an Assistant Professor in the computer science department at Michigan State University. He received a Ph.D. degree in Electrical and Computer Engineering program at Carnegie Mellon University in 2007. His research interests are in Computer Vision, Pattern Recognition and Machine Learning. He received the best paper award at the BTAS conference in 2013.



Anil Jain is a University Distinguished Professor in the Department of Computer Science at Michigan State University. His research interests include pattern recognition, computer vision and biometric authentication. He served as a member of the United States Defense Science Board and the Forensic Science Standards Board. He has received Fulbright, Guggenheim, Alexander von Humboldt, and IAPR King-Sun Fu awards and is a Fellow of ACM and IEEE. He is a member of the United States National Academy of Engineering and Foreign Fellow of the Indian National Academy of Engineering.

On the Capacity of Face Representations (Supplementary Material)

Sixue Gong, *Student Member, IEEE*, Vishnu Naresh Boddeti, *Member, IEEE*
and Anil K. Jain, *Life Fellow, IEEE*



In this supplementary material we include additional analysis and results that could not be included in the main paper due to space constraints.

1 GAUSSIAN CHANNEL

The validity of our capacity estimates are dependent on how well the face representations conform to the modeling assumptions of the Gaussian noise channel. The Gaussian noise channel is predicated upon the Gaussianity of the distributions P_y , P_x and P_z . While there are many tests to numerically quantify the Gaussianity of the embedding, most of these tests are either very unstable or are too conservative. Therefore, we resort to visual tests through Q-Q plots, univariate as well as multi-variate.

Univariate Test: Figure 1 displays the heat map of the pair-wise covariance between the dimensions of the representation averaged over the entire CASIA dataset. The covariance is predominantly a diagonal matrix with very low correlation between the feature dimensions. Figure 2 shows the marginal histograms and the normal distribution estimated from the data. Across all dimensions the data conforms very well to a normal distribution. This is also apparent in the marginal Q-Q plots for all the dimensions as shown in Fig. 3.

Multivariate Test: Figure 4 displays the Q-Q plot to compare the face representations of the entire CASIA dataset and the multi-variate Gaussian distribution estimated from the same representations. The plots show that apart from a few outlying samples the FaceNet representation does indeed follow a Gaussian distribution for both P_y and P_x .

2 Teacher-Student REPRESENTATIONS

The efficacy of our capacity estimation model is predicated on the ability of the *student* to mimic the *teacher*. It is critical that the *student* model preserves the overall structure of the embedding space. Figure 5 provides a visualization of the embeddings after projecting onto the top two principal eigenvectors of the embedding. We observe that the *student* model not only mimics the *teacher* at the level of each individual image but also preserves the overall global structure of the embedding space.

Our main motivation for learning a *student* model is to explicitly account for and extract uncertainty estimates

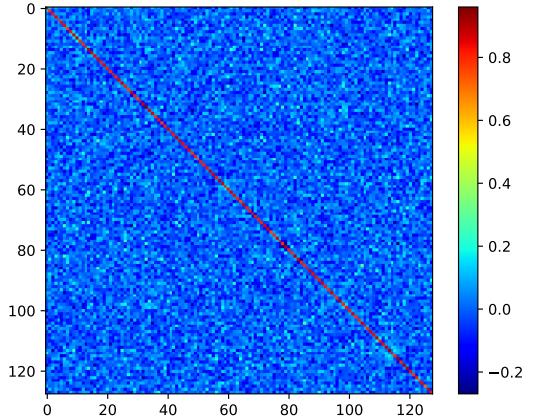


Fig. 1: Heat map of the empirical covariance of the representation estimated over the entire CASIA dataset.

TABLE 1: *Epistemic* and *Aleatoric* Uncertainty

Sample Class	<i>Epistemic</i>			<i>Aleatoric</i>		
	20	200	1000	20	200	1000
	3.30	3.47	3.54	7.90	7.96	7.99
	3.01	3.29	3.25	7.50	7.40	7.43
	7.05	7.04	7.13	13.01	13.02	13.14
	6.49	7.38	7.36	14.74	14.68	14.66

that the *teacher* itself is unable to provide. Furthermore, by explicitly accounting for the different sources of uncertainty, *epistemic* and *aleatoric*, the *student* model can identify *outlier* samples, those that are outside the distribution of the training data. For an *inlier* sample, the *epistemic* uncertainty can be explained away by the uncertainty in the model weights w while the same does not hold for an *outlier* image. In Fig. 5 we provide a pictorial demonstrative example of

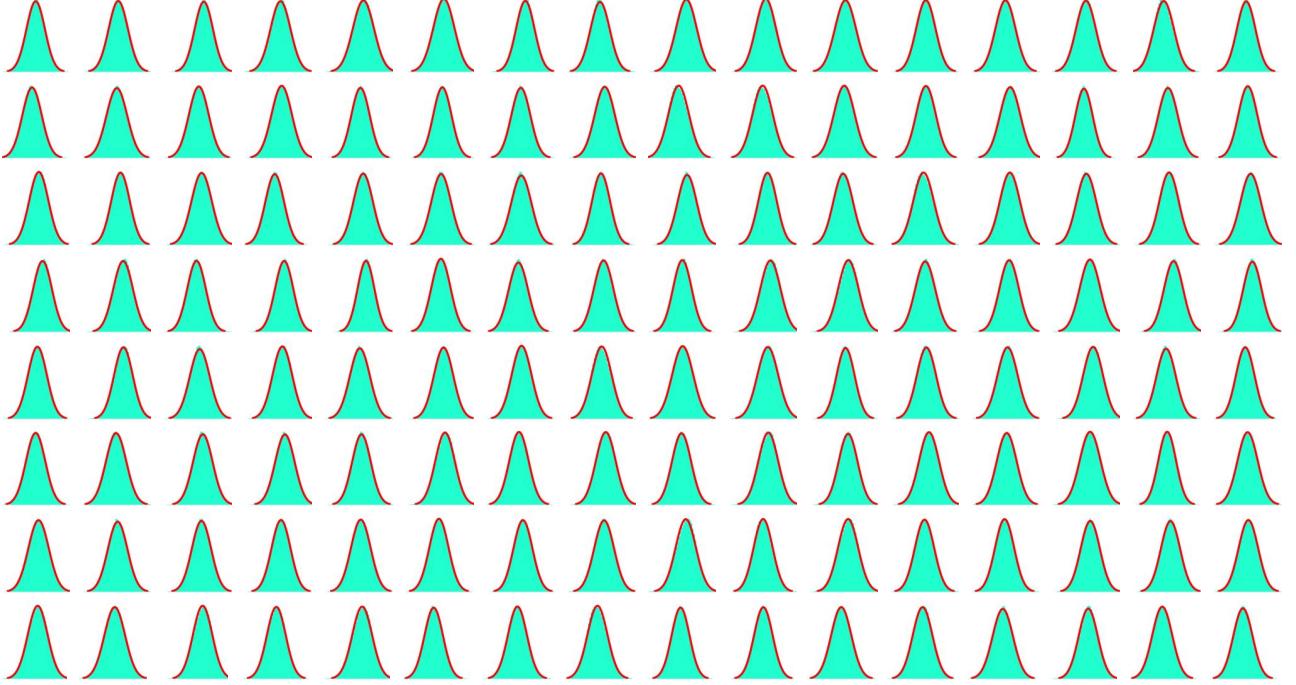


Fig. 2: Histogram and marginal distributions of the 128 dimensional FaceNet representation.

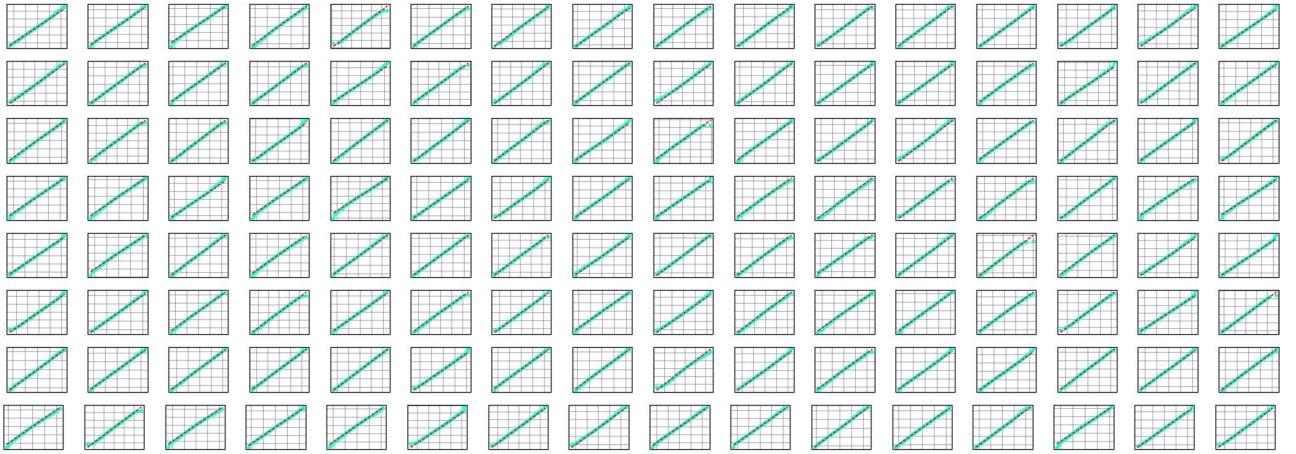


Fig. 3: Q-Q plot of the marginal distributions of the 128 dimensional FaceNet representation.

the embedding of an *outlier* and an *inlier* and quantitative examples in Table 1. Notice that, both qualitatively and quantitatively, the *outlier* example exhibits a greater amount of uncertainty compared to the *inlier* example.

3 CAPACITY ESTIMATES

Distribution of Class Uncertainty: The capacity estimates of a given representation is dependent on the choice of a canonical class-specific hyper-ellipsoid, that is representative of a typical subject among all possible subjects the face representation system is going to be used for. However, determining this canonical subject is a difficult task in itself. In Fig. 6 we show histograms of the distribution of class-specific hyper-ellipsoids for each of the datasets. Notice that classes in each dataset span a wide range of

hyper-ellipsoid sizes resulting in the wide range of capacity estimates. Some datasets even have a number of outlying classes with hyper-ellipsoids that are much bigger or much smaller than the rest of the dataset. These spurious classes (often the result of mis-labeled data) can have a significant effect on the capacity estimates. To overcome this problem, we manually identified and accounted for classes with mis-labeled images.

We also show examples of the classes and the images that correspond to the smallest and largest class hyper-ellipsoids in each dataset. First for FaceNet: PCSO in Fig. 7, LFW in Fig. 8, CASIA in Fig. 9, IJB-A in Fig. 10 and IJB-B in Fig. 11, followed by Eigenfaces: PCSO in Fig. 12, LFW in Fig. 13, CASIA in Fig. 14, IJB-A in Fig. 15 and IJB-B in Fig. 16.

Capacity at FAR: Figure 17 shows the capacity of Eigenfaces

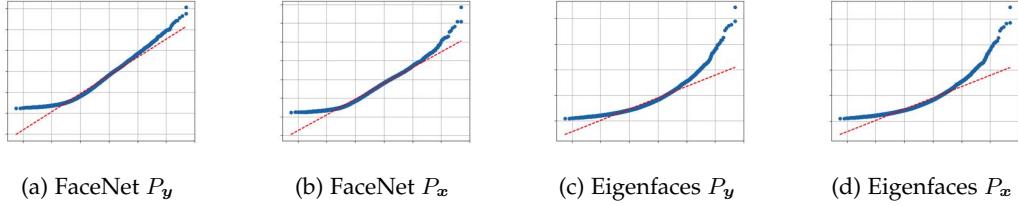


Fig. 4: Q-Q plot for each of the 128 dimensions of the FaceNet representation for both P_y and P_x

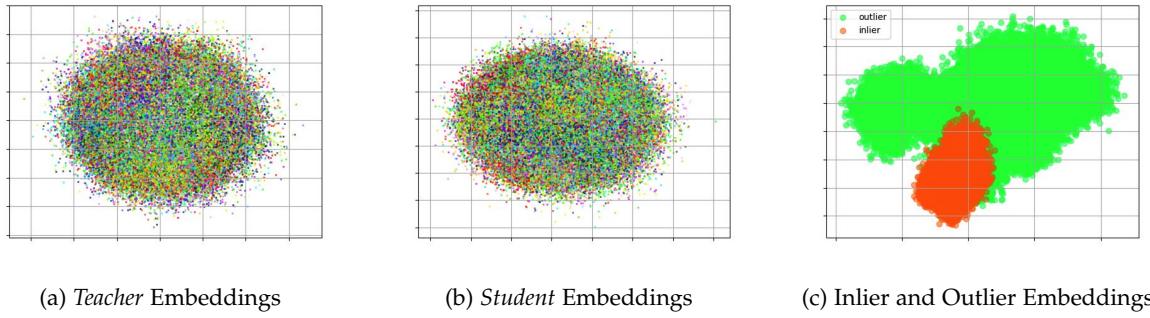


Fig. 5: Visualization of the embeddings of the (a) *teacher*, (b) *student*, (c) an inlier and outlier class. We project the face representation onto the first two principal eigenvectors for the purpose of this visualization.

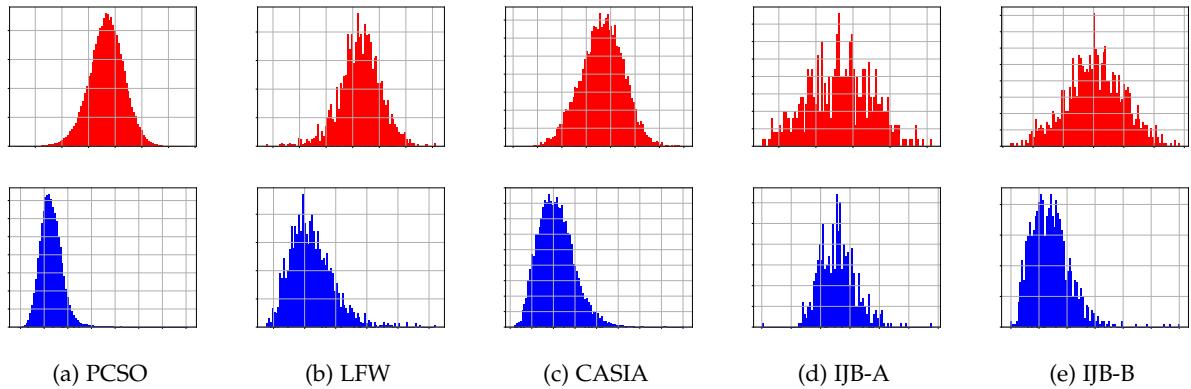


Fig. 6: Histogram of the class-specific hyper-ellipsoids for each dataset. Top-Row: FaceNet and Bottom-Row: Eigenfaces

on different datasets as a function of the operating point defined by the acceptable false accept rate.



(a) DNN: PCSO Large



(b) DNN: PCSO Small

Fig. 7: PCSO



(a) DNN: LFW Large



(b) DNN: LFW Small

Fig. 8: LFW



(a) DNN: CASIA Large



(b) DNN: CASIA Small

Fig. 9: CASIA



(a) DNN: IJB-A Large



(b) DNN: IJB-A Small

Fig. 10: IJB-A



(a) DNN: Large IJB-B



(b) DNN: Small IJB-B

Fig. 11: IJB-B

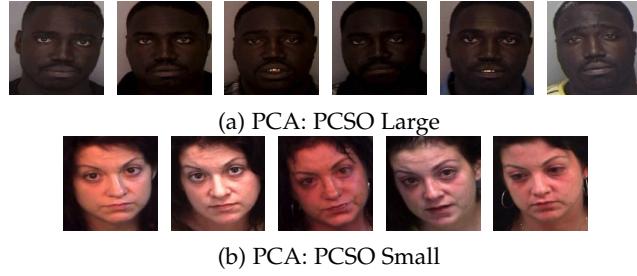


Fig. 12: PCSO



Fig. 13: LFW

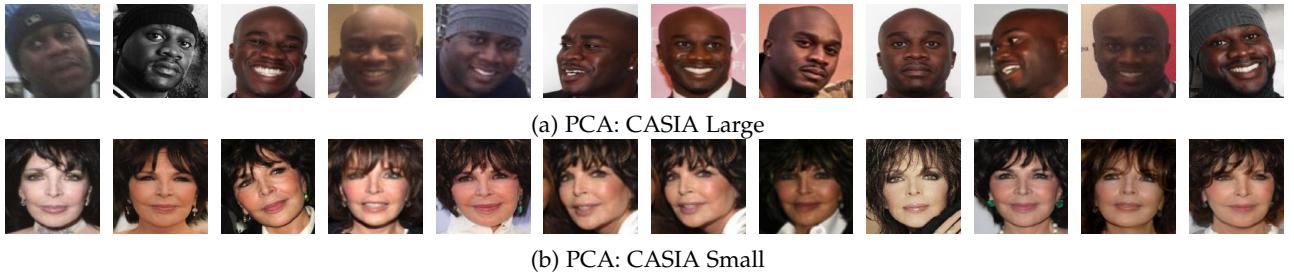


Fig. 14: CASIA

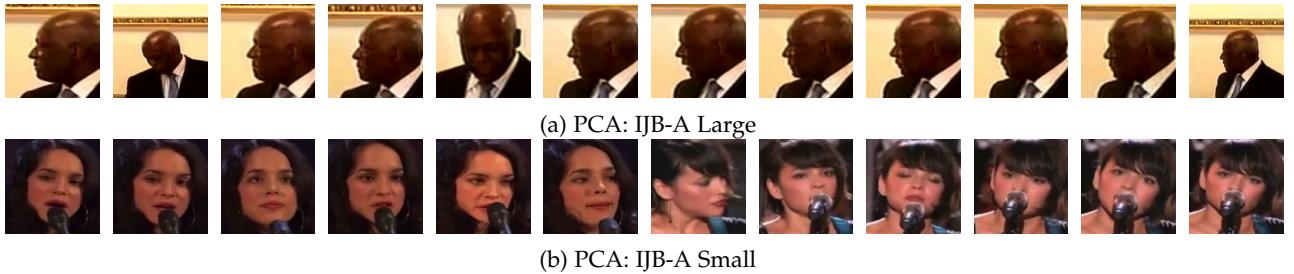


Fig. 15: IJB-A



Fig. 16: IJB-B

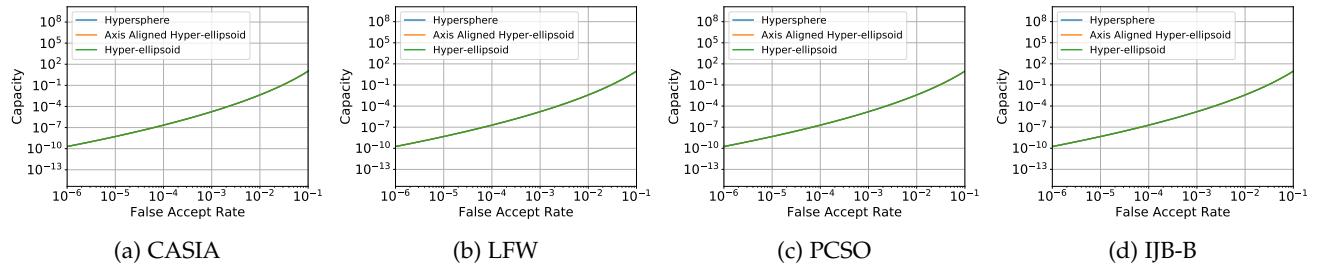


Fig. 17: For each of the datasets under consideration we estimate the capacity of Eigenfaces at different false accept rates corresponding to different operating points. Under the limit, the capacity tends to zero as the FAR tends to zero. Similarly the capacity tends to ∞ as the FAR tends to 1.0.