On the Intrinsic Dimensionality of Image Representations

Sixue Gong Vishnu Naresh Boddeti Anil K. Jain
Michigan State University, East Lansing MI 48824
{gongsixu, vishnu, jain}@msu.edu

Abstract

This paper addresses the following questions pertaining to the intrinsic dimensionality of any given image representation: (i) estimate its intrinsic dimensionality, (ii) develop a deep neural network based non-linear mapping, dubbed DeepMDS, that transforms the ambient representation to the minimal intrinsic space, and (iii) validate the veracity of the mapping through image matching in the intrinsic space. Experiments on benchmark image datasets (LFW, IJB-C and ImageNet-100) reveal that the intrinsic dimensionality of deep neural network representations is significantly lower than the dimensionality of the ambient features. For instance, SphereFace's [26] 512-dim face representation and ResNet's [16] 512-dim image representation have an intrinsic dimensionality of 16 and 19 respectively. Further, the DeepMDS mapping is able to obtain a representation of significantly lower dimensionality while maintaining discriminative ability to a large extent, 59.75% TAR @ 0.1% FAR in 16-dim vs 71.26% TAR in 512-dim on IJB-C [29] and a Top-1 accuracy of 77.0% at 19-dim vs 83.4% at 512-dim on ImageNet-100.

1. Introduction

An image representation is an embedding function that transforms the raw pixel representation of the image to a point in a high-dimensional vector space. Learning or estimating such a mapping is motivated by two goals: (a) the compactness of the representation, and (2) the effectiveness of the mapping for the task at hand. While the latter topic has received substantial attention, ranging from PCA based Eigenfaces [42] to deep neural network (DNN) based feature representations, there has been relatively little focus on the dimensionality of the representation itself. The dimensionality of image representations has ranged from hundreds to thousands of dimensions. For instance, current state-of-the-art image representations have 128, 512, 1024 and 4096 dimensions for FaceNet [35], ResNet [16], SphereFace [26] and VGG [36], respectively. The choice of dimensionality is often determined by practical considerations, such as, ease of learning the embedding function [38], constraints on system memory, etc. instead of the effective dimensionality necessary for image representation. This naturally raises the following fundamental but related questions, *How compact can the representation be without any loss in recognition performance?* In other words, *what is the intrinsic dimensionality of the representation?* And, *how can one obtain such a compact representation?* Addressing these questions is the primary goal of this paper.

The intrinsic dimensionality (ID) of a representation refers to the minimum number of parameters (or degrees of freedom) necessary to capture the entire information present in the representation [4]. Equivalently, it refers to the dimensionality of the m-dimensional manifold $\mathcal{M}$ embedded within the d-dimensional ambient (representation) space $\mathcal{P}$ where $m \leq d$. This notion of intrinsic dimensionality is notably different from common *linear dimensionality* estimates obtained through e.g., principal component analysis (PCA). This linear dimension corresponds to the best linear subspace necessary to retain a desired fraction of the variations in the data. In principle, linear dimensionality can be as large as the ambient dimension if the variation factors are highly entangled with each other. An illustration of these concepts is provided in Fig. 1.

The ability to estimate the intrinsic dimensionality of a given image representation is useful in a number of ways. At a **fundamental** level, the ID determines the true capacity and complexity of variations in the data captured by the representation, through the embedding function. In fact, the ID can be used to gauge the information content in the representation, due to its linear relation with Shannon entropy [41, 9]. Also, it provides an estimate of the amount of redundancy built into the representation which relates to its generalization capability. On a **practical** level, knowledge of the ID is crucial for devising optimal unsupervised strategies to obtain image features that are minimally redundant, while retaining its full ability to categorize images into different classes. Recognition in the intrinsic space can provide significant savings, both in memory requirements as well as processing time, across downstream tasks like large-scale face matching in the encrypted domain [5], im-

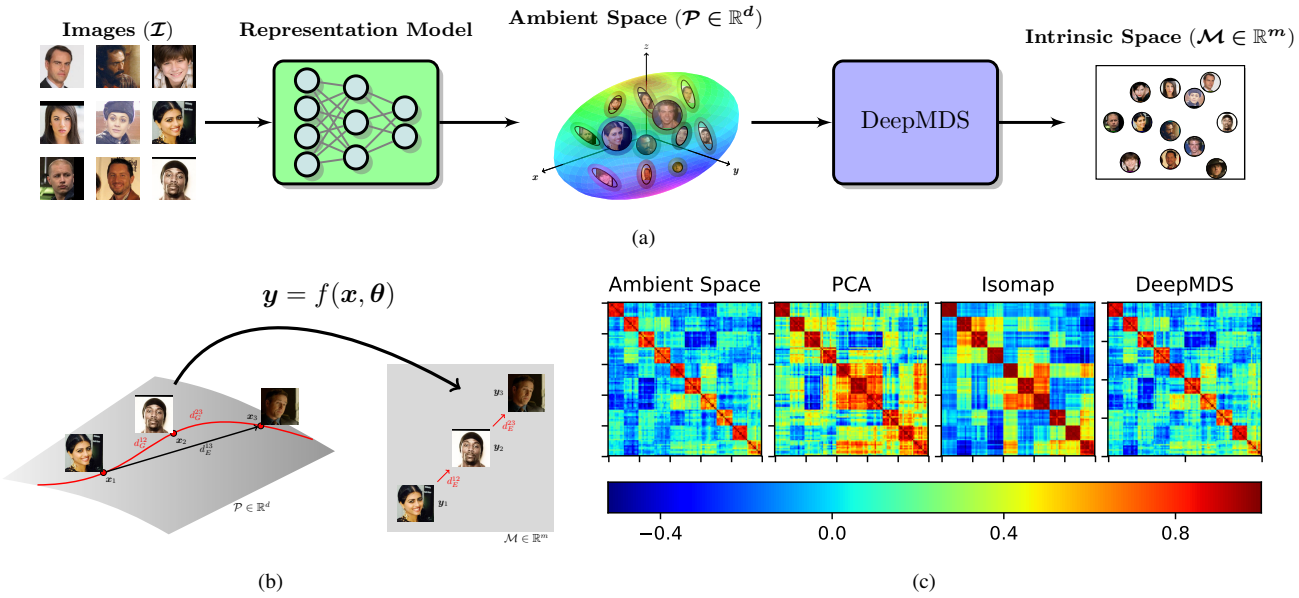


Figure 1: **Overview:** This paper studies the manifold of feature vectors of images $\mathcal{I}$ obtained from a given representation model. (a) We estimate the intrinsic dimensionality (ID) of the ambient space $\mathcal{P}$ and propose DeepMDS, an unsupervised method, to map $\mathcal{P}$ to a low-dimensional intrinsic space $\mathcal{M}$. (b) Illustration of the ambient space $\mathcal{P}$ and intrinsic manifold $\mathcal{M}$ of a face representation. Here, while the ambient and linear dimension of the representation is three, its ID is only two. (b) Heatmaps of similarity scores between face pairs of 10 classes with 10 images per class for a representation with ID of 10-dim. The similarity is computed in four different spaces, the 512-dim ambient space $\mathcal{P}$, 10-dim space of linear dimensionality (PCA), 10-dim intrinsic space $\mathcal{M}$ estimated by Isomap [40] and by our DeepMDS model. The class separability, as shown by the diagonal blocks, is better maintained by DeepMDS.

age matching and retrieval, etc. Lastly, gap between the ambient and intrinsic dimensionalities of a representation can serve as a useful indicator to drive the development of algorithms that can directly learn highly compact embeddings.

Estimating the ID of given data representation however is a challenging task. Such estimates are crucially dependent on the density variations in the representation, which in itself is difficult to estimate as images often lie on a topologically complex curved manifold [39]. More importantly, given an estimate of ID, how do we verify that it truly represents the dimensionality of the complex high-dimensional representation space? An indirect validation of the ID is possible through a mapping that transforms the ambient representation space to the intrinsic representation space while preserving its discriminative ability. However, there is no certainty that such a mapping can be found efficiently. In practice, finding such mappings can be considerably harder than estimating the ID itself.

We overcome both of these challenges by (1) adopting a topological dimensionality estimation technique based on the geodesic distance between points on the manifold, and (2) relying on the ability of DNNs to approximate the complex mapping function from the ambient space to the intrinsic space. The latter enables validation of the ID estimates through image matching experiments on the corresponding low-dimensional intrinsic representation of feature vectors.

The key contributions and findings of this paper are:

– The first attempt to estimate the intrinsic dimensionality of DNN based image representations.
– An unsupervised DNN based dimensionality reduction method under the framework of multidimensional scaling, called DeepMDS.
– Numerical experiments yield an ID estimate of, 12 and 16 for FaceNet [35] and SphereFace [26] face representations, respectively, and 19 for ResNet-34 [16] image representation. The estimates are significantly lower than their respective ambient dimensionalities, 128-dim for FaceNet and 512-dim for the others.
– DeepMDS mapping is significantly better than other dimensionality reduction approaches in terms of its discriminative capability.

2. Related Work

Image Representation: The quest to develop image representations that are simultaneously robust and discriminative have led to extensive research on this topic. Amongst the earliest learning based approaches, Turk and Pentland proposed Eigenfaces [42] that relied on principal component analysis (PCA) of data. Later on, integrated and high-dimensional spatially local features became prevalent for image recognition, notable examples include local binary patterns (LBP) [1], scale-invariant feature transform (SIFT) [28] and histogram of oriented gradients (HoG) [10]. In contrast to these hand-designed representations, the past

decade has witnessed the development of end-to-end representation learning systems. Convolutional neural network based features now typify the state-of-the-art image representations [16, 37, 26]. All of these representations are however characterized by features that range from hundreds to thousands of dimensions. While more compact representations are desirable, difficulties with optimizing DNNs with narrow bottlenecks [38] have proven to be the primary barrier towards realizing this goal.

Intrinsic Dimensionality: Existing approaches for estimating intrinsic dimensionality can be broadly classified into two groups: projection methods and geometric methods. The projection methods [11, 6, 43] determine the dimensionality by principal component analysis on local subregions of the data and estimating the number of dominant eigenvalues. These approaches have classically been used in the context of modeling facial appearance under different illumination conditions [12] and object recognition with varying pose [30]. While they serve as an efficient heuristic, they do not provide reliable estimates of intrinsic dimension. Geometric methods [31, 14, 7, 21, 17, 24] on the other hand model the intrinsic topological geometry of the data and are based on the assumption that the volume of a m-dimensional set scales with its size ϵ as ϵ^m and hence the number of neighbors less than ϵ also behaves the same way. Our approach in this paper is based on the topological notion of correlation dimension [14, 7], the most popular type of fractal dimensions. The correlation dimension implicitly uses nearest-neighbor distance, typically based on the Euclidean distance. However, Granata et.al. [13] observe that leveraging the manifold structure of the data, in the form of geodesic distances induced by a neighborhood graph of the data, provides more realistic estimates of the ID. Building upon this observation we base our ID estimates on the geodesic distance between points. We believe that estimating the intrinsic dimensionality would serve as the first step towards understanding the bound on the minimal required dimensionality for representing images and aid in the development of novel algorithms that can achieve this limit.

Dimensionality Reduction: There is a tremendous body of work on the topic of estimating low-dimensional approximations of data manifolds lying in high-dimensional space. These include linear approaches such as Principal Component Analysis [20], Multidimensional Scaling (MDS) [23] and Laplacian Eigenmaps [2] and their corresponding non-linear spectral extensions, Locally Linear Embedding [32], Isomap [40] and Diffusion Maps [8]. Another class of dimensionality reduction algorithms leverage the ability of deep neural networks to learn complex non-linear mappings of data including deep autoencoders [18], denoising autoencoders [44, 45] and learning invariant mappings either with the contrastive loss [15] or with the triplet loss [35]. While the autoencoders can learn a compact representation of data, such a representation is not explicitly designed to retain discriminative ability. Both the contrastive loss and the triplet loss have a number of limitations; (1) require similarity and dissimilarity labels from some source and cannot be trained in a purely unsupervised setting, (2) require an additional hyper-parameter, maximum margin of separation, which is difficult to pre-determine, especially for an arbitrary representation, and (3) do not maintain the manifold structure in the low-dimensional space. In this paper, we too leverage DNNs to approximate the non-linear mapping from the ambient to the intrinsic space. However, we consider an unsupervised setting (i.e., no similarity or dissimilarity labels) and cast the learning problem within the framework of MDS i.e., preserving the ambient graph induced geodesic distance between points in the intrinsic space.

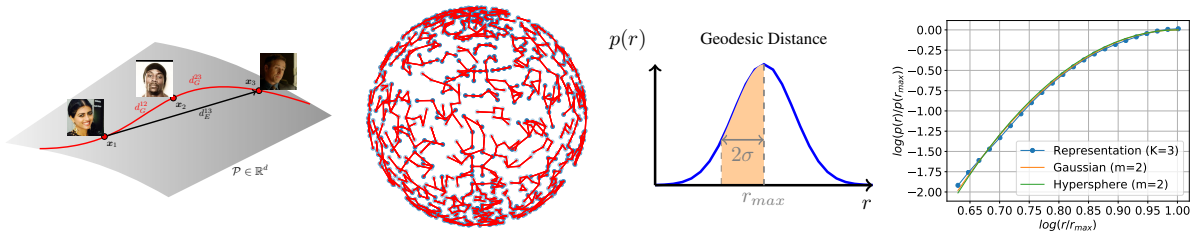
3. Approach

Our goal in this paper is to compress a given image representation space. We achieve this in two stages[1]: (1) estimate the intrinsic dimensionality of the ambient image representation, and (2) learn the DeepMDS model to map the ambient representation space $\mathcal{P} \in \mathbb{R}^d$ to the intrinsic representation space $\mathcal{M} \in \mathbb{R}^m$ ($m \leq d$). The ID estimates are based on the one presented by [13] which relies on two key ideas, (1) using graph induced geodesic distances to estimate the correlation dimension of the image representation topology, and (2) the similarity of the distribution of geodesic distances across different topological structures with the same intrinsic dimensionality. The DeepMDS model is optimized to preserve the interpoint geodesic distances between the feature vectors in the ambient and intrinsic space, and is trained in a stage-wise manner that progressively reduces the dimensionality of the representation. Basing the projection method on DNNs, instead of spectral approaches like Isomap, addresses the scalability and out-of-sample-extension problems suffered by spectral methods. Specifically, DeepMDS is trained in a stochastic fashion, which allows it to scale. Furthermore, once trained, DeepMDS provides a mapping function in the form of a feed-forward network that maps the ambient feature vector to its corresponding intrinsic feature vector. Such as map can easily be applied to new test data.

3.1. Estimating Intrinsic Dimension

We define the notion of intrinsic dimension through the classical concept of *topological dimension* of the support of a distribution. This is a generalization of the concept

[1]Traditional single-stage dimensionality reduction methods use visual aids to arrive at the final ID and intrinsic space, e.g., plotting the projection error against the ID values and looking for a "knee" in the curve.



(a) Graph Induced Geodesic Distance

(b) Topological Similarity

Figure 2: **Intrinsic Dimension:** Our approach is based on two observations: (a) Graph induced geodesic distance between images is able to capture the topology of the image representation manifold more reliably. As an illustration, we show the graph edges for the surface of a unitary hypersphere and a face manifold of ID two, embedded within a 3-dim space. (b) The distribution of the geodesic distances (for distance $r_{max} - 2\sigma \leq r \leq r_{max}$, where r_{max} is the distance at the mode) has been empirically observed [13] to be similar across different topological structures with the same intrinsic dimensionality. The plot shows the distance distribution for a face representation, unitary hypersphere and a Gaussian distribution of ID two embedded within 3-dim space.

of dimension of a linear space [2] to a non-linear manifold. Methods for estimating the topological dimension are all based on the assumption that the behavior of the number of neighbors of a given point on an m-dimensional manifold embedded within a d-dimensional space scales with its size ϵ as ϵ^m. In other words, the density of points within an ϵ-ball ($\epsilon \to 0$) in the ambient space is independent of the ambient dimension d and varies only according to its intrinsic dimensionality m. Given a collection of points $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$, the cumulative distribution of the pairwise distances $C(r)$ between the n points can be estimated as,

$$C(r) = \frac{2}{n(n-1)} \sum_{i<j=1}^{n} H(r - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|) = \int_0^r p(r)dr \qquad (1)$$

where $H(\cdot)$ is the Heaviside function and $p(r)$ is the probability distribution of the pairwise distances. In this paper, we choose the correlation dimension [14], a particular type of topological dimension, to represent the intrinsic dimension of the image representation. It is is defined as,

$$m = \lim_{r \to 0} \frac{\ln C(r)}{\ln r} \implies \lim_{r \to 0} C(r) \propto r^m \qquad (2)$$

Therefore, the intrinsic dimension is crucially dependent on the accuracy with which the probability distribution can be estimated at very small length-scales (distances), i.e., $r \to 0$. Significant efforts have been devoted to estimating the intrinsic dimension through line fitting in the $\ln C(r)$ vs $\ln r$ space around the region where $r \to 0$ i.e.,

$$m = \lim_{(r_2 - r_1) \to 0} \frac{\ln C(r_2) - \ln C(r_1)}{\ln r_2 - \ln r_1} \qquad (3)$$
$$= \lim_{r \to 0} \frac{d \ln C(r)}{d \ln r} = \lim_{r \to 0} \frac{p(r)}{C(r)} r = \lim_{r \to 0} m(r)$$

The main drawback with this approach is the need for reliable estimates of $p(r)$ at very small length scales, which is precisely where the estimates are most unreliable when data is limited, especially in very high-dimensional spaces. Granata et al. [13] present an elegant solution to this problem through three observations, (i) estimates of $m(r)$ can be stable even as $r \to 0$ if the distance between points is computed as the graph induced shortest path between points instead of the euclidean distance, as is commonly the case, (ii) the probability distribution $p(r)$ at intermediate length-scales around the mode of $p(r)$ i.e., $(r_{max} - 2\sigma) \leq r \leq r_{max}$ can be conveniently used to obtain reliable estimates of ID, and (iii) the distributions $p(r)$ of different topological geometries are similar to each other as long as the intrinsic dimensionality is the same, or in other words the distribution $p(r)$ depends only on the intrinsic dimensionality and not on the geometric support of the manifolds.

Figure 2 provides an illustration of these observations. Consider two different manifolds, faces and the surface of a $(m+1)$-dimensional unitary hypersphere (henceforth referred to as m-hypersphere $\mathcal{S}^m$), with intrinsic dimensionality of $m = 2$ but embedded within 3-dim Euclidean space. Beyond the nearest neighbor, the distance r between any pair of points in the manifold is computed as the shortest path between the points as induced by the graph connecting all the points in the representation. Figure 2b shows the distribution of $\log \frac{p(r)}{p(r_{max})}$ vs $\log \frac{r}{r_{max}}$ in the range $r_{max} - 2\sigma \leq r \leq r_{max}$, where σ is the standard deviation of $p(r)$ and $r_{max} = \arg\max_r p(r)$ corresponds to the radius of the mode of $p(r)$. Interestingly, different topological geometries, namely, a face representation of ID two, a 2-hypersphere and a 2-dim Gaussian, all embedded within 3-dim Euclidean space have almost identical distributions. More generally, the distribution of $\log \frac{p(r)}{p(r_{max})}$ vs $\log \frac{r}{r_{max}}$ in the range $r_{max} - 2\sigma \leq r \leq r_{max}$ is empirically observed to depend only on the intrinsic dimensionality, rather than the geometrical support of the manifold.

The intrinsic dimensionality of the representation mani-

[2] Linear dimension is the minimum number of independent vectors necessary to represent any given point in this space as a linear combination.

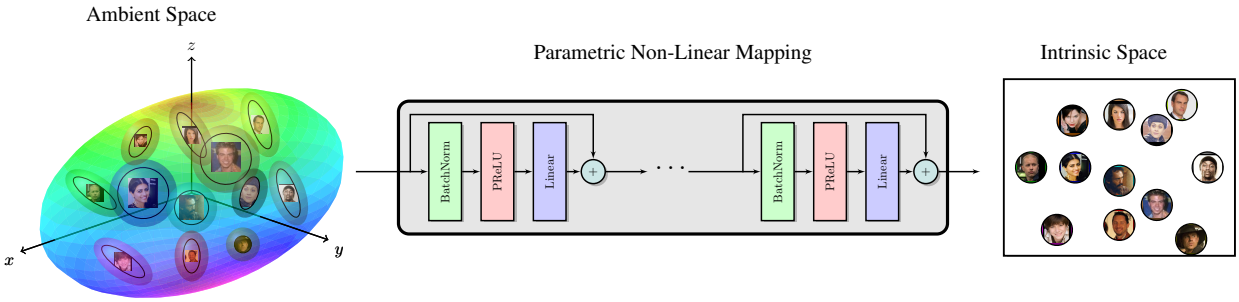


Figure 3: **DeepMDS Mapping:** A DNN based non-linear mapping is learned to transform the ambient space to a plausible intrinsic space. The network is optimized to preserve distances between pairs of points in the ambient and intrinsic space.

fold can thus be estimated by comparing the empirical distribution of the pairwise distances $\hat{p}_{\mathcal{M}}(r)$ on the manifold to that of a known distribution, such as the m-hypersphere in the range $r_{max} - \sigma \leq r \leq r_{max}$ (see supplementary material for Gaussian example). The distribution of the geodesic distance $p_{\mathcal{S}^m}(r)$ of m-hypersphere can be analytically expressed as, $p_{\mathcal{S}^m}(r) = c\sin^{m-1}(r)$, where c is a constant and m is the ID. Given $\hat{p}_{\mathcal{M}}(r)$, we minimize the Root Mean Squared Error (RMSE) between the distributions as,

$$\min_{c,m} \int_{r_{max}-2\sigma}^{r_{max}} \|\log \hat{p}_{\mathcal{M}}(r) - \log(c) - (m-1)\log\left(\sin[r]\right)\|^2$$

which upon simplification yields,

$$\min_{m} \int_{r_{max}-2\sigma}^{r_{max}} \left\| \log \frac{\hat{p}_{\mathcal{M}}(r)}{\hat{p}_{\mathcal{M}}(r_{max})} - (m-1)\log\left(\sin\left[\frac{\pi r}{2r_{max}}\right]\right)\right\|^2$$

The above optimization problem can be solved via a least-squares fit after estimating the standard deviation, σ, of $p(r)$ (see supplementary for details). Such a procedure could, in principle, result in a fractional estimate of dimension. If one only requires integer solutions, the optimal value of m can be estimated by rounding-off the least squares fit solution.

3.2. Estimating Intrinsic Space

The intrinsic dimensionality estimates obtained in the previous subsection alludes to the existence of a mapping, that can transform the ambient representation to the intrinsic space, but does not provide any solutions to find said mapping. The mapping itself could potentially be very complex and our goal of estimating it is practically challenging.

We base our solution to estimate a mapping from the ambient to the intrinsic space on Multidimensional scaling (MDS) [23], a classical mapping technique that attempts to preserve the distances (similarities) between points after embedding them in a low-dimensional space. Given data points $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ in the ambient space and $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ the corresponding points in the intrinsic low-dimensional space, the MDS problem is formulated as,

$$\min \sum_{i<j} \left(d_H(\boldsymbol{x}_i, \boldsymbol{x}_j) - d_L(\boldsymbol{y}_i, \boldsymbol{y}_j)\right)^2 \tag{4}$$

where $d_H(\cdot)$ and $d_L(\cdot)$ are distance (similarity) metrics in the ambient and intrinsic space, respectively. Different choices of the metric, leads to different dimensionality reduction algorithms. For instance, classical metric MDS is based on Euclidean distance between the points while using the geodesic distance induced by a neighborhood graph leads to Isomap [40]. Similarly, many different distance metrics have been proposed corresponding to non-linear mappings between the ambient space and the intrinsic space. A majority of these approaches are based on spectral decompositions and suffer many drawbacks, (i) computational complexity scales as $\mathcal{O}(n^3)$ for n data points, (ii) ambiguity in the choice of the correct non-linear function, and (iii) collapsed embeddings on more complex data [15].

To overcome these limitations, we employ a DNN to approximate the non-linear mapping that transforms the ambient representation, $\boldsymbol{x}$, to the intrinsic space, $\boldsymbol{y}$ by a parametric function $\boldsymbol{y} = f(\boldsymbol{x}; \boldsymbol{\theta})$ with parameters $\boldsymbol{\theta}$. We learn the parameters of the mapping within the MDS framework,

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n}\sum_{i=1}^{n} \left[d_H(\boldsymbol{x}_i, \boldsymbol{x}_j) - d_L(f(\boldsymbol{x}_i;\boldsymbol{\theta}), f(\boldsymbol{x}_j;\boldsymbol{\theta}))\right]^2 + \lambda\|\boldsymbol{\theta}\|_2^2$$

where the second term is a regularizer with a hyperparameter λ. Figure 3 shows an illustration of the DNN based mapping.

In practice, directly learning the mapping from the ambient to the intrinsic space is very challenging, especially for disentangling a complex manifold under high levels of compression. We adopt a curriculum learning [3] approach to overcome this challenge and progressively reduce the dimensionality of the mapping in multiple stages. We start with easier sub-tasks and progressively increase the difficulty of the tasks. For example, a direct mapping from $\mathbb{R}^{512} \rightarrow \mathbb{R}^{15}$ is instead decomposed into multiple mapping functions $\mathbb{R}^{512} \rightarrow \mathbb{R}^{256} \rightarrow \mathbb{R}^{128} \rightarrow \mathbb{R}^{64} \rightarrow \mathbb{R}^{32} \rightarrow \mathbb{R}^{15}$. We formulate the learning problem for L mapping functions $\left(\boldsymbol{y}^l = f_l(\boldsymbol{x};\boldsymbol{\theta})\right)$ as:

$$\min_{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_L} \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{L} \alpha_l \left[d_H(\boldsymbol{x}_i, \boldsymbol{x}_j) - d_L(\boldsymbol{y}_i^l, \boldsymbol{y}_j^l)\right]^2 + \lambda\|\boldsymbol{\theta}_l\|_2^2$$

where $\boldsymbol{\theta}_l$ are the parameters of the l-th mapping. Appropriately scheduling the α_l weights enables us to set it up as a curriculum learning problem.

4. Experiments

In this section, first we will estimate the intrinsic dimensionality of multiple image representations over multiple datasets of varying complexity. Then, we will evaluate the efficacy of the proposed DeepMDS model in finding the mapping from the ambient to the intrinsic space while maintaining its discriminative ability.

4.1. Datasets

We choose two different domains of classification problems for our experiments, face verification and image classification. We consider two different face datasets for the former and the ImageNet ILSVRC-2012 for the latter. Recall that DeepMDS is an unsupervised method, so category information associated with the objects or faces is neither used for intrinsic dimensionality estimation nor for learning the mapping from the ambient to intrinsic space.

LFW [19]: 13,233 face images of 5,749 subjects, downloaded from the web. These images 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.

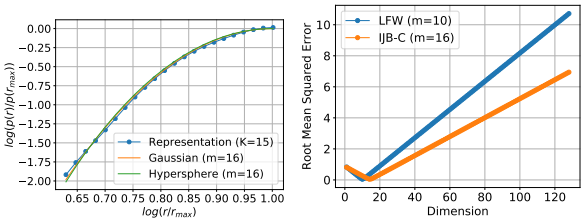
IJB-C [29]: IARPA Janus Benchmark-C (IJB-C) dataset consists of 3,531 subjects with a total of 31,334 (21,294 face and 10,040 non-face) still images and 11,779 videos (117,542 frames), an average of 39 images per subject. This dataset emphasizes faces with full pose variations, occlusions and diversity of subject occupation and geographic origin. Images in this dataset are labeled with ground truth bounding boxes and other covariate meta-data such as occlusions, facial hair and skin tone.

ImageNet [34]: The ImageNet ILSVRC-2012 classification dataset consists of 1000 classes, with 1.28 million images for training and 50K images for validation. We use a subset of this dataset by randomly selecting 100 classes with the largest number of images, for a total of 130K training images and 5K testing images.

4.2. Representation Models

For the face-verification task, we consider multiple publicly available state-of-the-art face embedding models, namely, 128-dim FaceNet [35] representation and 512-dim SphereFace [26] representation. In addition, we also evaluate a 512-dim variant of FaceNet[3] that outperforms the 128-dim version. All of these representations are learned from the CASIA WebFace [47] dataset, consisting of 494,414 images across 10,575 subjects. For image classification on the ImageNet dataset, we choose a pre-trained 34 layer version of the ResNet [16] architecture.

[3]https://github.com/davidsandberg/facenet



(a) Distance Distribution $p(r)$ (b) Least Squares Fitting

Figure 4: **Intrinsic Dimensionality:** (a) Geodesic distance distribution, and (b) global minimum of RMSE.

4.3. Baseline Methods

Intrinsic Dimensionality: We select two different algorithms for estimating the intrinsic dimensionality of a given representation, a classical k-nearest neighbor based estimator [31] and "Intrinsic Dimensionality Estimation Algorithm" (IDEA) [33].

Dimensionality Reduction: We compare DeepMDS against three dimensionality reduction algorithms, principal component analysis (PCA) for linear dimensionality reduction, Isomap [40] and denoising autoencoders [45] (DAE).

4.4. Intrinsic Dimensions

Implementation Details: The ID estimates for all the methods we evaluate are dependent on the number of neighbors k. For the baselines, k is used to compute the parameters of the probability density. For our method, k parameterizes the construction of the neighborhood graph. For the latter, the choice of k is constrained by three factors; (1) k should be small enough to avoid shortcuts between points that are close to each other in the Euclidean space, but are potentially far away in the corresponding intrinsic manifold due to highly complicated local curvatures. (2) On the other hand, k should also be large enough to result in a connected graph i.e., there are no isolated data samples., and (3) k that best matches the geodesic distance distribution of a hypersphere of the same ID i.e., k that minimizes the RMSE. Figure 4a shows the distance distributions for SphereFace with $k = 15$, a 16-hypersphere and a 16-dim Gaussian. The close similarity of the pairwise distance distributions of these manifolds in the graph induced geodesic distance space suggests that the ID of SphereFace (512-dim ambient space) is 16. Figure 4b shows the optimal RMSE for SphereFace[4] at different values of m. For all the approaches we select the k-nearest neighbors using cosine similarity for SphereFace, Euclidean distance for ResNet and arc-length,

[4]Similar curves for other representations and datasets can be found in the supplementary material.

Table 1: Intrinsic Dimensionality: Graph Distance [13]

Representation	dataset	k			
		4	7	9	15
FaceNet-128	LFW	**10***	13	11	18
	IJB-C	10	10	10	**11***
FaceNet-512	LFW	**10***	11	11	17
	IJB-C	11	11	12	**12***
SphereFace	LFW	**10***	11	13	9
	IJB-C	14	14	16	**16***
ResNet-34	ImageNet-100	16	18	**19***	23

Table 2: LFW Face Verification for SphereFace Embedding

Dimension	Dimension Reduction method			
	PCA	Isomap	DAE	DeepMDS
512	96.74%			
256	**96.75**%	92.88%	77.80%	96.73%
128	**96.80**%	93.18%	32.95%	96.44%
64	91.71%	95.00%	32.04%	**96.50**%
32	66.38%	95.31%	11.71%	**96.31**%
16	32.67%	89.47%	27.53%	**95.95**%
10 (ID)	16.04%	77.31%	6.73%	**92.33**%

$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \cos^{-1}\left(\frac{\boldsymbol{x}_1^T \boldsymbol{x}_2}{\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|}\right)$, for FaceNet features, as the latter are normalized to reside on the surface of a unitary hypersphere. Finally, for simplicity, we round the ID estimates to the nearest integer for all the methods.

Experimental Results: Table 1 reports the ID estimates from the graph method for different values of k[5] and for different representation models across different datasets. Due to lack of space we report the ID estimates of the baselines in the supplementary material. We make a number of observations from our results: (1) Surprisingly, the ID estimates across all the datasets, feature representations and ID methods are significantly lower than the dimensionality of the ambient space, between 10 and 20, suggesting that image representations could, in principle, be almost $10\times$ to $50\times$ more compact. (2) Both[6] the k-NN based estimator [31] and the IDEA estimator [33] are less sensitive to the number of nearest neighbors in comparison to the graph distance based method [13], but are known to underestimate ID for sets with high intrinsic dimensionality [43].

4.5. Dimensionality Reduction

Given the estimates of the dimensionality of the intrinsic space, we learn the mapping from the ambient space to *a plausible* intrinsic space with the goal of retaining the discriminative ability of the representation. The true intrinsic representation (ID and space) is unknown and therefore not feasible to validate directly. However, verifying its discriminate power can serve to indirectly validate both the ID estimate and the learned intrinsic space.

Implementation Details: We first extract image features through the representations i.e., FaceNet-128, FaceNet-512 and SphereFace for face images and ResNet-34 for ImageNet-100. The architecture of the proposed DeepMDS model is based on the idea of skip connection laden residual units [16]. We train the mapping from the ambient to intrinsic space in multiple stages with each stage comprising of two residual units. Once the individual stages are trained, all the L projection models are jointly fine-tuned to maintain the pairwise distances in the intrinsic space. We adopt a similar network structure (residual units) and training strategy (stagewise training and fine-tuning) for the stacked denoising autoencoder baseline. From an optimization perspective, training the autoencoder is more computationally efficient than the DeepMDS model, $\mathcal{O}(n)$ vs $\mathcal{O}(n^2)$.

The parameters of the network are learned using the Adam [22] optimizer with a learning rate of 3×10^{-4} and the regularization parameter $\lambda = 3 \times 10^{-4}$. We observed that using the cosine-annealing scheduler [27] was critical to learning an effective mapping. To facilitate classification on ImageNet in the intrinsic space, after learning the projection, we separately learn a linear as well as a k-nearest neighbor (k-NN) classifier on the projected feature vectors of the training set.

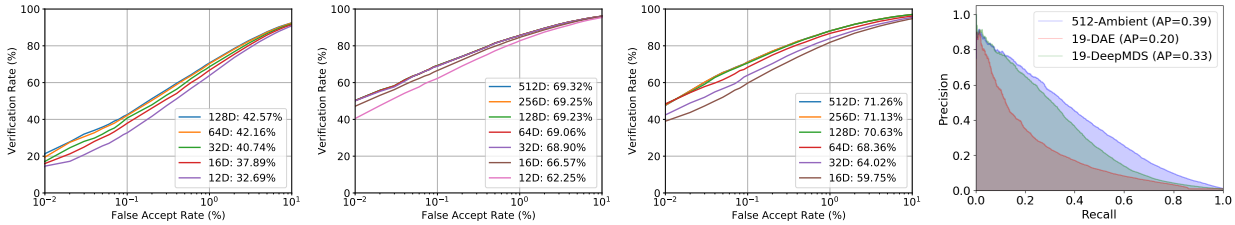
Experimental Results: We evaluate the efficacy of the learned projections, namely PCA, Isomap and DeepMDS, in the learned intrinsic space and compare their respective performance in the ambient space. Face representations are evaluated in terms of verification (TAR @ FAR) performance and classification on ImageNet-100 in terms of accuracy (Top-1 and Top-5). Given the ID estimate, designing an appropriate scheme for mapping the intrinsic manifold is much more challenging than the ID estimation itself. To show how dimensionality of the intrinsic space influences the performance of image representations, we evaluate and compare their performance at multiple intermediate spaces.

Face verification is performed on the IJB-C dataset following its verification protocol and on the LFW dataset following the BLUFR [25] protocol. Due to space constraints we only show results on the DeepMDS model here, corresponding results for the baseline dimensionality reduction methods can be found in the supplementary material. Figure 5 shows the ROC curves for the IJB-C dataset and the precision-recall curves for a image retrieval task on ImageNet-100. Table 2 reports the verification rate at FAR of 0.1% on the LFW dataset. Similarly, Table 3 shows the Top-1 and Top-5 accuracy on ImageNet-100 for a pretrained ResNet-34 representation via a parametric (linear) as well as a non-parametric (k-NN) classifier.

We make the following observations from these results: (1) for all the tasks the performance of the DeepMDS features up to 32 dimensions (for faces) is comparable to the original 128-dim and 512-dim features. The 10-dim space

[5] * denotes final ID estimate that satisfies all constraints on k.

[6] Reported in supplementary material due to space constraints.



(a) FaceNet-128 (b) FaceNet-512 (c) SphereFace (d) Image Retrieval

Figure 5: Face Verification on IJB-C [29] (TAR @ 0.1% FAR in legend) for the (a) FaceNet-128, (b) FaceNet-512 and (c) SphereFace embeddings and (d) Image retrieval on ImageNet-100 for the ambient 512-dim ResNet-34 representation, the intrinsic 19-dim space obtained from DAE and DeepMDS.

Table 3: ImageNet-100 Classification (%) for ResNet-34

	Classifier	Method	Dimension					
			512	256	128	64	32	19 (ID)
Top-1	Linear	DAE	80.0	**80.9**	73.2	70.0	63.1	50.2
		DeepMDS	80.0	79.4	**76.1**	**71.4**	**70.2**	**68.0**
	k-NN	DAE	83.4	**81.3**	**79.1**	76.4	76.7	73.4
		DeepMDS	83.4	80.9	78.7	**77.8**	**77.1**	**77.0**
Top-5	Linear	DAE	96.0	**95.5**	90.2	**88.0**	84.2	76.5
		DeepMDS	96.0	95.3	**93.1**	85.2	**85.2**	**84.8**

Table 4: DeepMDS Training Methods (TAR @ 0.1% FAR)

Method	Direct	Direct+IS	Stagewise + Finetune	Stagewise
TAR	80.25	86.15	90.42	**92.33**

of DeepMDS on LFW, consisting largely of frontal face images with minimal pose variations and facial occlusions, achieves a TAR of 92.33% at 0.1% FAR, a loss of about 4.5% compared to the ambient space. The 12-dim space of DeepMDS on IJB-C, with full pose variations, occlusions and diveristy of subject, achieves a TAR of 62.25% at 0.1% FAR, compared to 69.32% in the ambient space. (2) the proposed DeepMDS model is able to learn a low-dimensional space up to the ID with a performance penalty of 5%-10% for compression factors of 30$\times$ to 40$\times$ for 512-dim representations, underscoring the fact that learning a mapping from ambient to intrinsic space is more challenging than estimating the ID itself. (3) In both tasks, we observe that the DeepMDS model is able to retain significantly more discriminative ability compared to the baseline approaches even at high levels of compression. Although DAE achieves comparative results on ImageNet-100 classification, DeepMDS significantly outperforms DAE for image retrieval tasks. While Isomap is more competitive than the other baselines it suffers from some drawbacks: (i) Due to its iterative nature, it does not provide an explicit mapping function for new (unseen) data samples, while the autoencoder and DeepMDS models can map such data samples. Therefore, Isomap cannot be utilized to evaluate classification accuracy on the validation/test set of ImageNet-100 dataset, and (ii) Computational complexity of Isomap is $\mathcal{O}(n^3)$ and hence does not scale well to large datasets (IJB-C, ImageNet) and needs approximations, such as Nyström approximation [39], for tractability.

Ablation Study: Here we demonstrate the efficacy of the stagewise learning process for training the DeepMDS model. All models have the same capacity. We consider four variants: (1) **Direct** mapping from the ambient to intrinsic space, (2) **Direct+IS:** direct mapping from ambient to intrinsic space with intermediate supervision at each stage i.e., optimize aggregate intermediate losses, (3) **Stagewise** learning of the mapping, and (4) **Stagewise+Fine-Tune:** the projection model trained stagewise and then fine-tuned. Table 4 compares the results of these variations on the LFW dataset (BLUFR protocol). Our results suggest that stagewise learning of the non-linear projection models is more effective at progressively disentangling the ambient representation. Similar trend was observed on larger datasets (IJB-C and ImageNet). In fact, stagewise training with fine-tuning was critical in learning an effective projection, both for DeepMDS as well as DAE.

5. Concluding Remarks

This paper addressed two questions, given a DNN based image representation, what is the minimum degrees of freedom in the representation i.e., its intrinsic dimension and can we find a mapping between the ambient and intrinsic space while maintaining the discriminative capability of the representation? Contributions of the paper include, (i) a graph induced geodesic distance based approach to estimate the intrinsic dimension, and (ii) DeepMDS, a non-linear projection to transform the ambient space to the intrinsic space. Experiments on multiple DNN based image representations yielded ID estimates of 9 to 20, which are significantly lower than the ambient dimension (10$\times$ to 40 $\times$). The DeepMDS model was able to learn a projection from ambient to the intrinsic space while preserving its discriminative ability, to a large extent, on the LFW, IJB-C and ImageNet-100 datasets. Our findings in this paper suggest that image representations could be significantly more compact and call for the development of algorithms that can directly learn more compact image representations.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. *European Conference on Computer Vision*, 2004. 2

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. 3

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *International Conference on Machine Learning*, pages 41–48. ACM, 2009. 5

[4] R. S. Bennett. Representation and analysis of signals part xxi. the intrinsic dimensionality of signal collections. Technical report, Johns Hopkins University Baltimore MD, Deptartment of Electrical Engineering and Computer Science, 1965. 1

[5] V. N. Boddeti. Secure face matching using fully homomorphic encryption. In *IEEE International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, 2018. 1

[6] J. Bruske and G. Sommer. Intrinsic dimensionality estimation with optimally topology preserving maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):572–575, 1998. 3

[7] F. Camastra and A. Vinciarelli. Estimating the intrinsic dimension of data with a fractal-based method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(10):1404–1407, 2002. 3

[8] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. 3

[9] J. A. Costa and A. O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8):2210–2221, 2004. 1

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 2

[11] K. Fukunaga and D. R. Olsen. An algorithm for finding intrinsic dimensionality of data. *IEEE Transactions on Computers*, 100(2):176–183, 1971. 3

[12] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001. 3

[13] D. Granata and V. Carnevale. Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets. *Scientific Reports*, 6:31377, 2016. 3, 4, 7

[14] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. In *The Theory of Chaotic Attractors*, pages 170–189. Springer, 2004. 3, 4

[15] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1735–1742, 2006. 3, 5

[16] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016. 1, 2, 3, 6, 7

[17] M. Hein and J.-Y. Audibert. Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$. In *International Conference on Machine Learning*, 2005. 3

[18] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. 3

[19] 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, Technical Report 07-49, University of Massachusetts, Amherst, 2007. 6

[20] I. T. Jolliffe. Principal component analysis and factor analysis. In *Principal Component Analysis*, pages 115–128. Springer, 1986. 3

[21] B. Kégl. Intrinsic dimension estimation using packing numbers. In *Advances in Neural Information Processing Systems*, 2003. 3

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[23] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964. 3, 5

[24] E. Levina and P. J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, 2005. 3

[25] S. Liao, Z. Lei, D. Yi, and S. Z. Li. A benchmark study of large-scale unconstrained face recognition. In *IEEE International Joint Conference on Biometrics (IJCB)*, 2014. 7

[26] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3, 6

[27] I. Loshchilov and F. Hutter. SGDR: stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 7

[28] D. G. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision*, 1999. 2

[29] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, et al. Iarpa janus benchmark–c: Face dataset and protocol. In *International Conference on Biometrics*, 2018. 1, 6, 8

[30] H. Murase and S. K. Nayar. Visual learning and recognition of 3-d objects from appearance. *International Journal of Computer Vision*, 14(1):5–24, 1995. 3

[31] K. W. Pettis, T. A. Bailey, A. K. Jain, and R. C. Dubes. An intrinsic dimensionality estimator from near-neighbor information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):25–37, 1979. 3, 6, 7

[32] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. 3

[33] A. Rozza, G. Lombardi, C. Ceruti, E. Casiraghi, and P. Campadelli. Novel high intrinsic dimensionality estimators. *Machine Learning*, 89(1-2):37–65, 2012. 6, 7

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 6

[35] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1, 2, 3, 6

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[37] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, volume 4, page 12, 2017. 3

[38] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Web-scale training for face identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 1, 3

[39] A. Talwalkar, S. Kumar, and H. Rowley. Large-scale manifold learning. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008. 2, 8

[40] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 2, 3, 5, 6

[41] J. Theiler. Estimating fractal dimension. *JOSA A*, 7(6):1055–1073, 1990. 1

[42] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1991. 1, 2

[43] P. J. Verveer and R. P. W. Duin. An evaluation of intrinsic dimensionality estimators. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):81–86, 1995. 3, 7

[44] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, pages 1096–1103. ACM, 2008. 3

[45] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010. 3, 6

[46] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 6

[47] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. 6