Gaussian Affinity for Max-margin Class Imbalanced Learning

Munawar Hayat*†   Salman Khan*‡   Syed Waqas Zamir*   Jianbing Shen**   Ling Shao*

*Inception Institute of Artificial Intelligence, †University of Canberra,
‡Australian National University, *Beijing Institute of Technology

firstname.lastname@inceptioniai.org

Abstract

Real-world object classes appear in imbalanced ratios. This poses a significant challenge for classifiers which get biased towards frequent classes. We hypothesize that improving the generalization capability of a classifier should improve learning on imbalanced datasets. Here, we introduce the first hybrid loss function that jointly performs classification and clustering in a single formulation. Our approach is based on an 'affinity measure' in Euclidean space that leads to the following benefits: (1) direct enforcement of maximum margin constraints on classification boundaries, (2) a tractable way to ensure uniformly spaced and equidistant cluster centers, (3) flexibility to learn multiple class prototypes to support diversity and discriminability in feature space. Our extensive experiments demonstrate the significant performance improvements on multiple imbalanced datasets belonging to visual classification and verification tasks. The proposed loss can easily be plugged in any deep architecture as a differentiable block and demonstrates robustness against different levels of data imbalance and corrupted labels.

1. Introduction

Deep neural networks are data hungry in nature and require large amounts of data for successful training. For imbalanced datasets, where several (potentially important) classes have a scarce representation, the learned models are biased towards highly abundant classes. This is because the scarce classes have less representations during training which results in a mismatch between the joint distribution model for training $p(x, y)$ and test sets $p(x', y')$. This leads to lower recall rates for rare classes, which are otherwise critically desirable in numerous scenarios. As an example, a malignant lesion is rare compared to benign ones, but should not be miss-classified.

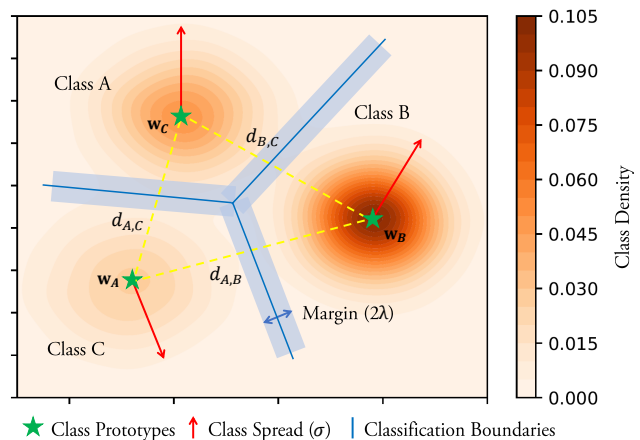The first two authors contributed equally

Figure 1. Affinity Loss integrates classification and clustering in a single objective. It's flexible formulation in Euclidean space allows enforcing margin between classes, control over learned clusters, number of class-prototypes and the distance between class-prototypes. Such max-margin learning greatly helps in overcoming class imbalance by learning balanced classification regions and generalizable class boundaries.

The soft-max loss is a popular choice for conventional recognition tasks. However, through extensive experiments, we show that it is less suitable to handle mismatch between train and test distributions. This is partly due to no direct enforcement of margins in the classification space and the lack of a principled approach to control intra-class variations and inter-class separation. Here, we propose that max-margin learning can improve generalization which can help mitigate classifier bias towards more frequent classes by learning balanced representations for all classes. Remarkably, some recent efforts focus on introducing max-margin constraints within the soft-max loss function [10, 34, 33]. Since soft-max loss computes similarities in the angular domain (vector dot-product or cosine similarity), direct enforcement of angular margins is ill-posed and existing works either involve approximations or make restricting assumptions (e.g., points lying on a hypersphere).

This paper proposes a novel loss formulation that en-

hances generalization by jointly reducing intra-class variations and maximizing inter-class distances. A notable difference from the previous works is the automatic learning of class representative prototypes in the Euclidean space with inherent flexibility to enforce certain geometric constraints. This is in contrast to soft-max loss where more abundant classes tend to occupy additional space in the projected feature space and rare classes get a skewed representation. The proposed objective is named the 'Affinity loss function' as it is based on a Gaussian similarity metric defined in terms of Bergman divergence. The proposed loss formulation learns to map input images to a discriminative Euclidean space where the distance with class representative prototypes provides a direct similarity measure for each class. The class prototypes are key points in the embedding space around which feature points are clustered [46].

The affinity loss function promotes the classifier to have a simpler, balanced and more generalizable inductive bias during training. The proposed loss function thus provides the following advantages: (1) An inherent mechanism to jointly cluster and classify feature vectors in the Euclidean space, (2) A tractable way to ensure uniformly spaced and equidistant class prototypes (when embedding dimension d and prototype number n are related as: $n < d + 1$), (3) Along-with uniformly spaced prototypes, our formulation ensures that the clusters formed around the prototypes are uniformly shaped (in terms of second order moments), and (4) The resulting classifier shows robustness against different levels of label noises and imbalances amongst classes.

The proposed loss function is a differentiable module which is applicable to different network architectures, and complements the commonly deployed regularization techniques including dropout, weight decay and momentum. Through extensive evaluations on a number of datasets, we demonstrate that it achieves a highly balanced and generalizable classifier, leading to significant improvements over previous techniques.

2. Related Work

Class-imbalanced Learning: Imbalanced datasets exhibit complex characteristics and learning from such data requires designing new techniques and paradigms. The existing class imbalance approaches can be divided into two main categories, **1)** data-level, and **2)** algorithm-level approaches. The data-level schemes modify the distribution of data e.g., by oversampling the minority classes [44, 7, 14, 15, 21] or undersampling the majority classes [26, 3]. Such approaches are usually susceptible to redundancy and over-fitting (for over-sampling) and critical information loss (for under-sampling). In comparison, the algorithm level approaches improve the classifier itself e.g., through cost-sensitive learning. Such methods incorporate prior knowledge about classes based upon their significance

or representation in the training data [27, 41, 24]. These methods have been applied to different classifiers including SVMs [51], decision trees [64] and boosting [52]. Some works further explore ensemble of cost-sensitive classifiers to tackle imbalance [19, 25]. A major challenge associated with these cost-sensitive methods is that the class-specific costs are only defined at the beginning, and they lack mechanisms to dynamically update the costs during the course of training.

Deep Imbalanced Learning: Some recent attempts have been made to learn deep models from imbalanced data [20, 24, 5, 55, 39, 22]. For example, the method in [20] first learns to under sample the training data using a neural network, followed by Synthetic Minority Oversampling TEchnique (SMOTE) based technique to re-balance the data. Deep models are trained to directly optimize the imbalanced classification accuracy in [55, 39]. Wang et al. [56] propose a meta learning approach to progressively transfer the model parameters from majority towards less-frequent classes. Khan et al. [22] used uncertainty estimates to tackle class imbalance problem. Some works [24, 5] train cost sensitive deep networks, which alternatively optimize class costs and network weights. Continually determining class costs while training a deep model is still an open and challenging research problem, and makes optimization intractable in learning from large scale datasets [18].

Joint Loss Formulation: Popular loss functions used for classification in deep networks include hinge loss, soft-max loss, Euclidean loss and contrastive loss [23]. A triplet loss could simultaneously perform recognition and clustering, however its training is prohibitive due to huge number of triplet combinations on large-scale datasets [43]. Since these loss functions are limited in their capability to achieve discriminability in feature space, recent literature explores the combination of multiple loss functions [38]. To this end, [47] showed that the combination of soft-max and contrastive losses concurrently enforce intra-class compactness and inter-class separability. On a similar line, [57] proposed 'center loss' that uses separate objectives for classification and clustering.

Max-margin Learning: Margin-maximizing learning objectives have been traditionally used in machine learning. Hinge loss in Support vector machines is one of the pioneering max-margin learning framework [16]. Some recent works aim to integrate max-margin learning with cross-entropy loss function. Among these, Large-margin soft-max [34] enforces inter-class separability directly on the dot-product similarity while SphereFace [33] and ArcFace [10] enforce multiplicative and additive angular margins on the hypersphere manifold, respectively. The hypersphere assumption for feature space makes the resulting loss less generalizable to applications other than face recognition. Furthermore, enforcing margin based separation in angu-

lar domain is an ill-posed problem and either requires approximations or assumptions (e.g., unit sphere) [12]. This paper proposes a new flexible loss function which simultaneously performs clustering and classification, and enables direct enforcement of the max-margin constraints. We describe the proposed loss formulation next.

3. Max-margin Framework

We propose a hybrid multi-task formulation to perform learning on imbalanced datasets. The proposed formulation combines classification and clustering in a single objective that minimizes intra-class variations while simultaneously achieving maximal inter-class separation. We first explain why traditional Soft-max Loss (SL) is unsuitable for large-margin learning and then introduce our novel objective function.

3.1. Soft-max Loss

Given an input-output pair $\{\mathbf{x}_i, y_i\}$, a deep neural network transforms input to a feature space representation $\mathbf{f}_i$ using a function $\mathcal{F}$ parameterized by θ i.e., $\mathbf{f} = \mathcal{F}(\mathbf{x}; \theta)$. The soft-max loss can then compute the discrepancy between prediction and ground-truth in the label space as follows:

$$L_{sm} = \frac{1}{N} \sum_i -\log\Big(\frac{\exp(\mathbf{w}_{y_i}^T \mathbf{f}_i)}{\sum_j \exp(\mathbf{w}_j^T \mathbf{f}_i)} \Big), \qquad (1)$$

where $i \in [1, N]$, $j \in [1, C]$, N and C are number of training examples and classes respectively. It is worth noting that we have included the last fully connected layer in the definition of soft-max loss which will be useful in further analysis. Also, for the sake of brevity, we do not mention unit biases in Eq. 1.

Although soft-max loss is one of the most popular choices for multi-class classification, in the following discussion, we argue that it is not suitable for class imbalanced learning due to several limitations.

Limitations of SL: The loss function in Eq. 1 computes inner vector product $\langle \mathbf{w}, \mathbf{f} \rangle$ which measures the projection of feature representation on to each of the class vectors $\mathbf{w}_j$. The goal is to perfectly align $\mathbf{f}_i$ with the correct class vector $\mathbf{w}_{y_i}$ such that the data likelihood is maximized. Due to the reliance of the soft-max loss on vector dot-product, it has the following limitations: (1) No inherent mechanism to ensure max margin constraints. Computation of inter-class margin for soft-max loss is intractable [12]. Large margin constraints promote better generalization in imbalanced distributions and robustness against input perturbations [9], (2) The learned projection vectors are not necessarily equispaced in the classification space. That is, ideally the angle between closest projection vectors should be equal (e.g., $\frac{2\pi}{k}$

in 2D where k is the number of classes). However, in practice the projection vectors for majority classes occupy more angular space compared with minority classes. This has been visualized in Fig. 2 on imbalanced MNIST dataset, and leads to poor generalization to test samples, and (3) The length $\|\mathbf{w}_j\|_2$ of the learnt projection vectors for different classes is not necessarily the same. It has been shown in the literature that the minority class projection vectors are weaker (i.e., with less magnitude) compared with the majority classes [34]. Cost-sensitive learning which artificially augments the magnitude of the minority class projection vectors has been shown to be effective for imbalance learning [24].

Unsuitability of SL for Imbalanced Learning: We attribute the above limitations to not directly enforcing the max-margin constraints on the classification boundaries. Considering the definition of soft-max loss (Eq. 1) in terms of dot-products $\mathbf{w}^T \mathbf{f}$, we can simplify the expression as follows:

$$L_{sm}^i \propto \sum_{j \neq y_i} \exp(\mathbf{w}_j^T \mathbf{f}_i - \mathbf{w}_{y_i}^T \mathbf{f}_i) \qquad (2)$$

The decision boundary for a class pair $\{j, k\}$ is given by the case where $\mathbf{w}_j^T \mathcal{F}(\mathbf{x}) = \mathbf{w}_k^T \mathcal{F}(\mathbf{x})$, i.e., the class boundaries are shared between the pair of classes. Further, minimization of L_{sm}^i requires $\mathbf{w}_j^T \mathcal{F}(\mathbf{x}) > \mathbf{w}_k^T \mathcal{F}(\mathbf{x}) : k \neq j$ for correct class assignment to $\mathbf{x}$. This is a '*relative constraint*' and therefore the soft-max loss L_{sm} does not necessarily: (a) reduces intra-class variations, (b) enforces a margin between each class pair. To address these issues, we propose our new loss function next.

3.2. Max-margin Learning with Gaussian Affinity

Euclidean space similarity measure: Instead of computing similarities with class prototypes using vector dot-product, we propose to measure class similarities for an input feature in the Euclidean space using a Gaussian similarity measure in terms of Bergman divergence (squared ℓ^2 distance):

$$d(\mathbf{f}_i, \mathbf{w}_j) = \exp\Big(-\frac{\|\mathbf{f}_i - \mathbf{w}_j\|^2}{\sigma} \Big), \qquad (3)$$

where, σ denotes a weighting parameter. This provides us: **(a)** the flexibility to directly enforce margin maximizing constraints, **(b)** have equi-spaced classification boundaries for multiple classes, **(c)** control the variance of learned clusters and therefore enhancing intra-class compactness, **(d)** the freedom to use standard distance measures in Euclidean domain to measure similarity and most importantly **(e)** simultaneous classification and clustering in a single objective function. We note that such distance based similarity functions have been used for semi-supervised [35] or
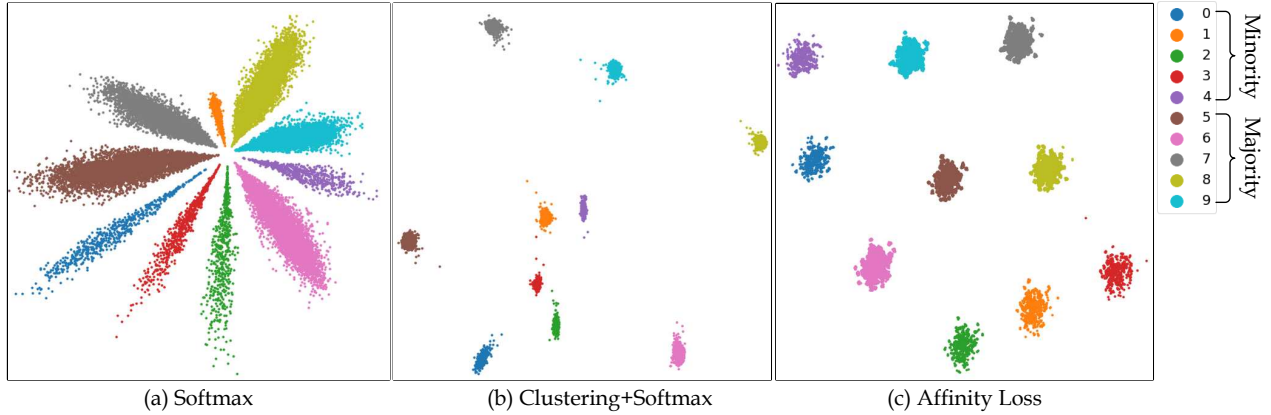
Figure 2. 2D feature space projections in terms of penultimate layer activations (i.e., considering only 2 neurons for 2D visualization purposes). The model is trained on imbalanced MNIST data (by retaining only 10% of the samples for digits 0-4) using different losses: (a) soft-max loss learns floral petals in angular space, note that the minority class feature vectors are weaker (shorter in length) and occupy less angular space, (b) center loss reduces intra-class variations by performing clustering. However, the minority class vectors tend to be congested near the center and are confused amongst each other, (c) the proposed affinity loss learns equi-spaced clusters of uniform shapes for both majority and minority classes.

few-shot settings [46], but have not been explored before for class-imbalanced learning in a max-margin framework. The function in Eq. 3 is a valid metric as we show below.

Proposition 1. *The similarity function $d(\mathbf{a}, \mathbf{b})$ is a valid similarity metric for any real-valued inputs.*

Proof. The real-valued similarity function $d(\mathbf{a}, \mathbf{b})$ will define a valid similarity metric if it satisfies the following conditions [31]:

- Non-negativity: $d(\mathbf{a}, \mathbf{b}) \geq 0$
- Symmetry: $d(\mathbf{a}, \mathbf{b}) = d(\mathbf{b}, \mathbf{a})$
- Equivalence: $d(\mathbf{a}, \mathbf{a}) = d(\mathbf{b}, \mathbf{b}) = d(\mathbf{a}, \mathbf{b})$ iff $\mathbf{a} = \mathbf{b}$
- Self-similarity: $d(\mathbf{a}, \mathbf{a}) \geq d(\mathbf{a}, \mathbf{b})$
- Triangular similarity: $d(\mathbf{a}, \mathbf{b}) + d(\mathbf{b}, \mathbf{c}) \leq d(\mathbf{a}, \mathbf{c}) + d(\mathbf{b}, \mathbf{b})$.

Since, all above conditions are true for $d(\cdot)$, therefore, it is a valid similarity metric. $\square$

Relation between Dot-product and Gaussian Similarity: The proposed Gaussian similarity measure is related to the dot-product as follows:

$$d(\mathbf{f}_i, \mathbf{w}_j) = \exp\left(-\frac{\|\mathbf{f}_i\|^2 + \|\mathbf{w}_j\|^2 - 2\langle \mathbf{w}_j, \mathbf{f}\rangle}{\sigma}\right), \quad (4)$$

$$\langle \mathbf{w}_j, \mathbf{f}\rangle = \frac{\sigma \log d(\mathbf{f}_i, \mathbf{w}_j) + \|\mathbf{f}_i\|^2 + \|\mathbf{w}_j\|^2}{2} \quad (5)$$

Intuitively, the above relation implies the dependence of soft-max loss on the scale/magnitude of feature vectors

and class prototypes. It leads to two conclusios: (1) It can be seen that $d(\mathbf{f}_i, \mathbf{w}_j)$ is bounded between $[0, 1]$ since $\|\mathbf{f}_i\|^2 + \|\mathbf{w}_j\|^2 \geq 2\langle \mathbf{w}_j, \mathbf{f}\rangle$, while $\langle \mathbf{w}_j, \mathbf{f}\rangle$ can have large magnitudes. (2) The Gaussian measure can be considered as an inverse chord distance when magnitudes of vectors are normalized to be equal. The dot-product in that case is directly proportional to the Gaussian similarity and both similarity measures will behave similarly if no additional constraints are included in our proposed similarity measure. However, the main flexibility with our formulation is the explicit introduction of margin constraints, which we introduce next.

Enforcing margin between classes: Note that some variants of soft-max loss introduce angle based margin constraints [33, 10], however, the margins in angular domain are computationally expensive and implemented only as approximations due to intractability. Our formulation allows a more direct margin penalty in the loss function. The proposed max margin loss function based on Eq. 3 is given by,

$$L_{mm} = \sum_j \max\left(0, \lambda + d(\mathbf{f}_i, \mathbf{w}_j) - d(\mathbf{f}_i, \mathbf{w}_{y_i})\right) \quad (6)$$

where $j \neq y_i$, $d(\mathbf{f}_i, \mathbf{w}_{y_i})$ is the similarity of the sample with its true class, $d(\mathbf{f}_i, \mathbf{w}_j)$ is its similarity with other classes, and λ is the enforced margin.

Uniform classification regions: The soft-max loss does not ensure uniform classification regions for all classes. As a result, under-sampled minority classes get a shrinked representation in the feature space compared to more frequent classes. To ensure equi-distant weight vectors, we propose
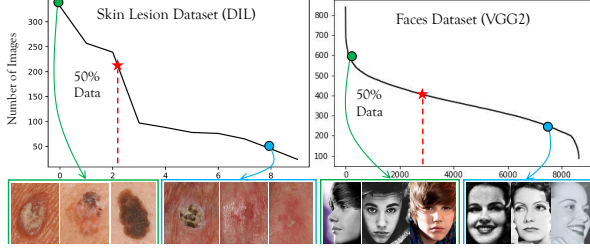
Figure 3. Data Imbalance due to long-tail distribution.

to apply a regularization on the learned class weights. This regularizer is termed as a '*diversity regularizer*' as it enforces all class centers (**w**) to be uniformly spread out in the feature space. The diversity regularizer is formally defined as follows:

$$R(\mathbf{w}) = \mathbb{E}[(\|\mathbf{w}_j - \mathbf{w}_k\|^2 - \mu)^2], \ s.t. \ j < k, \quad (7)$$

$$\mu = \frac{2}{C^2 - C} \sum_{j < k} \|\mathbf{w}_j - \mathbf{w}_k\|^2, \quad (8)$$

where μ is the mean distance between all class prototypes. The pair-wise Euclidean distances above can be efficiently computed using vectorization and tensor slicing through quadratic expansion.

Multi-centered learning: For challenging classification problems, the feature space may be partitioned such that all samples belonging to the same class are not co-located in a single region. Therefore, clustering all same class samples with a single prototype (class center) will not be optimal in such cases. To resolve this limitation, we introduce a novel multi-centered learning paradigm based on our max-margin framework. Instead of learning a single projection vector $\mathbf{w}_j$ for each class, the proposed framework enables learning multiple projection vectors $\{\mathbf{w}_t\}_j$ per-class. Specifically, we can learn m projection vectors per class, where similarity of a feature vector $\mathbf{f}_i$ with a class j is given by:

$$d(\mathbf{f}_i, \mathbf{w}_j) = \max \left\{ \exp \left(- \frac{\|\mathbf{f}_i - \mathbf{w}_{j,t}\|^2}{\sigma} \right) \right\}, t = [1, m]. \quad (9)$$

Affinity loss is then defined similar to Eq. 6 above. The overall loss function therefore becomes:

$$L = L_{mm} + R(\mathbf{w}). \quad (10)$$

The diversity regularizer for the multi-center case is enforced on the similarity between all $m * C$ prototypes.

4. Experiments

To demonstrate the effectiveness of the proposed affinity loss, we perform experiments on datasets which exhibit natural imbalance. These include Dermofit Image Library

(DIL) for skin lesion classification and large scale image datasets for facial verification. We further extensively evaluate various components of the proposed approach by systematically generating imbalance and introducing different levels of label noise. Through these empirical evaluations, we provide an evidence of the robustness of the proposed method against different data imbalance levels and noisy training labels. A brief description about the evaluated datasets is presented next.

4.1. Datasets

Skin Melanoma Dataset (DIL): Edinburgh Dermofit Image Library (DIL) contains 1300 images belonging to 10 skin lesion categories including melanomas, seborrhoeic keratosis and basal cell carcinomas. The images are based upon diagnosis from dermatologists and dermatopathologists. The number of images vary amongst categories (between 24 and 331, mean 130, median 83), and show significant imbalance, with 50% of all images belonging to only top two classes (Fig. 3). Similar to [2], we perform two experiments, considering five and ten class splits respectively, and report results for 3-fold cross validation.

Face Recognition: Datasets used to train large scale face recognition models have natural imbalance. This is because the data is web-crawled, and images for some identities are easily available in abundance compared with others. For unconstrained face recognition, we train our model on VGG2 [4], which is a large scale dataset with inherent class imbalance. We evaluate the trained network on four different datasets. These include two popular widely used benchmarks i.e., Labelled Faces in the Wild (LFW) [28] and YouTube Faces (YTF) [58]. We further evaluate on Celebrities in Frontal Profile (CFP) [42] and Age Database (AgeDB) [37].

VGG2: facial image dataset [4] contains 3.31 million images belonging to 8, 631 identities. The number of samples for each subject exhibit imbalance and vary from 80 to 843 with a mean of 362. The data is collected from the Internet and has real-life variations in the form of ethnicities, head poses, illumination changes and age groups.

LFW: Labelled Faces in the Wild (LFW) [28] contains 13, 233 static images of 5749 individuals collected over the Internet in real-life situations. We follow the standard evaluation protocol 'unrestricted with labeled outside data' [28] and test on 6000 pairs for face verification.

YTF: YouTube Faces (YTF) [58] has 3425 videos belonging to 195 different subjects. The length of video sequences varies between 48 and 6070 frames, with an average of 181.3 frames. We follow the standard evaluation protocol for face verification on 5000 video pairs.

CFP: contains 10 frontal and 4 profile view images for 500 identities [42]. Two evaluation protocols are used based upon the type of images in the gallery and probe: frontal-

frontal (FF) and frontal-profile (FP). Each protocol has 10 runs, each with 700 face pairs (350 same and 350 different). **AgeDB:** has $12,240$ images acquired in-the-wild for 440 subjects [37]. Along with variations across expression deformations, head poses and illumination conditions, a distinct feature of this dataset is the diversity across ages which ranges between 3 and 101 years, with an average of 49 years. Test data has four groups with different age gaps (5, 10, 20 and 30 years). Each group contains ten splits, each having 600 face image pairs (300 same, 300 different). We use the most challenging split with 30 years gap.

Imbalanced MNIST: Standard MNIST has $70,000$ handwritten images of digits (0-9), $60,000$ of these images are used for training (600/class) and the remaining 10,000 for testing (100/class). For this paper, we perform experiments on the standard evaluation split, as well as by systematically creating imbalance in the training set. For this, we reduce the even and odd digit samples to 10% and 25%. We further perform ablative study (Sec. 4.6) by gradually introducing different imbalance ratios amongst classes and noise levels in the training labels.

4.2. Experimental Settings

For experiments on DIL dataset, ResNet-18 backbone is used in combination with the proposed affinity loss. For training the model to learn features for face verification tasks, we deploy Squeeze and Excitation (SE) networks [17] with ResNet-50 backbone and affinity loss. The face images are cropped and re-sized to 112×112 using multi-task cascaded Convolution Neural Network (CNN) [62]. The model is trained using random horizontal flips as data augmentation. The features extracted after the global pooling layer are then used for face verification evaluations on different datasets. The experiments on MNIST are performed on a simple network with four hidden layers having three convolution layers (32, 64 and 128 filters of 5×5), one fully connected layer (128 neurons), and an output layer. The model is trained with Stochastic Gradient Descent (SGD) optimizer with momentum and learning rate decay. For ablative study in Sec. 4.6, we only change the output soft-max layer with the proposed Affinity loss layer and keep rest of the architecture fixed.

4.3. Results and Analysis

Table 2 present our experimental results on DIL dataset. In Exp#1, we report average performance for 3 fold cross validation on five classes (Actinic Keratosis, Basal Cell Carcinoma, Melanocytic Nevus, Squamous Cell Carcinoma and Seborrhoeic Keratosis). Compared with existing state of the art [24], we achieve an absolute gain of 10.9% on Exp#1. For Exp#2 on DIL dataset, all 10 classes are considered. Evaluations on 3 fold cross validation in Table 2 show a significant performance improvement of 7.7% for

Methods (using stand. split)	Performances	
Deeply Supervised Nets [30]	99.6%	
Generalized Pooling Func. [29]	99.7%	
Maxout NIN [6]	**99.8%**	
Imbalanced ($\downarrow$)	CoSen CNN [24]	Affinity Loss
Stand. split	99.3%	**99.6%**
10% of odd digits	98.6%	**99.3%**
10% of even digits	98.4%	**99.3%**
25% of odd digits	98.9%	**99.4%**
25% of even digits	98.5%	**99.5%**

Table 1. Evaluations on Imbalanced MNIST Database.

Methods (using stand. split)	Performances	
	Exp#1 (5-classes)	Exp#2 (10-classes)
Hierarchical-KNN [2]	$74.3 \pm 2.5\%$	$68.8 \pm 2.0\%$
Hierarchical-Bayes [1]	$69.6 \pm 0.4\%$	$63.1 \pm 0.6\%$
Flat-KNN [2]	$69.8 \pm 1.6\%$	$64.0 \pm 1.3\%$
CoSen CNN [24]	$80.2 \pm 2.5\%$	$72.6 \pm 1.6\%$
Affinity Loss	$\mathbf{91.1 \pm 1.7\%}$	$\mathbf{80.3 \pm 2.1\%}$

Table 2. Evaluation on DIL Database.

Exp#2. Confusion matrix analysis for class-wise accuracy comparison in Fig 4 shows that the performance boost is more pronounced for minority classes with lower representations. We attribute this to the capability of the proposed method to simultaneously optimize within class compactness by performing feature space clustering, and enhance inter-class separability by enforcing max-margin constraints. Our method achieves competitive performance on LFW and YTF datasets in Table 3. The performances on LFW and YTF are already saturated with many recent methods surpassing human-level results. The top performing methods on these datasets have been trained on much larger models with significantly more data and model parameters. Further evaluations on other facial recognition benchmarks achieve verification accuracies of 95.9%, 99.5% and 96.0% on AgeDB30, CFP-FF and CFP-FP datasets respectively. These results prove the effectiveness of the proposed approach for large scale imbalanced learning. It is worth noting that our proposed Affinity loss does not require additional compute and memory and is easily scalable to larger datasets. This is in contrast to some of the existing loss formulations (such as triplet loss [43] and contrastive loss [13]) which do enhance feature space discriminability, but suffer scalability to large data due to substantial possible combinations of training pairs, triplets or quintuplets.

4.4. Generalization across Imbalance Levels

To test the generalization of the proposed method for different imbalance levels, we gradually reduce the training set by changing the representation of the minority class samples on MNIST data. Specifically, we gradually alter the majority to minority class ratios (up-to $1:0.025$) by ran-
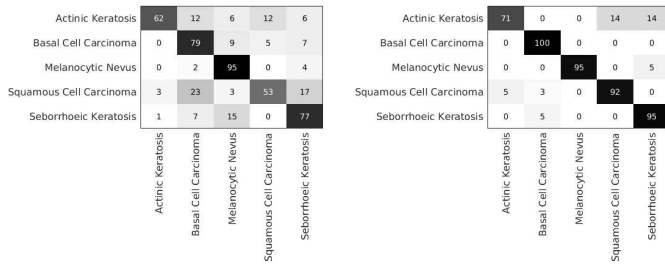
(a) CosSen CNN [24]   (b) Affinity Loss

Figure 4. Confusion matrices for Exp#1 on DIL dataset.

Methods	#Models	Train Data	LFW	YTF
DeepFace [49]	3	4M	97.35	91.4
FaceNet [43]	1	200M	99.63	95.4
Web-scale [50]	4	4.5M	98.37	-
VGG Face [40]	1	2.6M	98.95	97.3
DeepID2+ [48]	25	0.3M	99.47	93.2
Baidu [32]	1	1.3M	99.13	-
Center Face [57]	1	0.7M	99.28	94.9
Marginal Loss [11]	1	4M	99.48	95.98
Noisy Softmax [8]	1	Ext. WebFace	99.18	94.88
Range Loss [63]	1	1.5M	99.52	93.7
Augmentation [36]	1	WebFace	98.06	-
Center Invariant Loss [59]	1	WebFace	99.12	93.88
Feature transfer [61]	1	4.8M	99.37	-
Softmax+Contrastive [47]	1	WebFace	98.78	93.5
Triplet Loss [43]	1	WebFace	98.7	93.4
Large Margin Softmax [34]	1	WebFace	99.10	94.0
Center Loss [57]	1	WebFace	99.05	94.4
SphereFace [33]	1	WebFace	99.42	95.0
CosFace [54]	1	WebFace	99.33	96.1
LMLE [18]	1	WebFace	99.51	95.8
Affinity Loss	1	VGG2	**99.65**	**97.3**

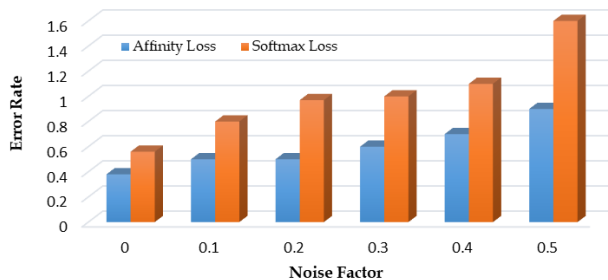Table 3. Face Verification Performance on LFW and YTF datasets.

Figure 5. The effect of label noise on the soft-max and affinity loss functions.

domly dropping samples of the first five digits $(0-4)$. Under these settings, we therefore have significantly lower representation for half of the classes. The experimental results in terms of error rates against fraction of retained minority class samples are shown in Fig. 6. We also repeat these experiments for standard soft-max loss. The comparison in Fig. 6 demonstrates a consistently superior performance of the proposed loss function across all settings. The effect on achieved performance is more noticeable for larger im-
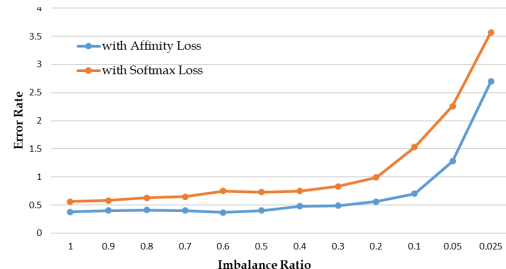

Figure 6. Robustness analysis against different imbalance levels (fraction of retained minority class samples)

balance levels between majority and minority classes. The proposed Affinity loss enhances inter-class separability irrespective of the class frequencies by enforcing margin maximization constraints. Soft-max loss does not have inherent margin learning characteristics. Further, compared with soft-max loss, where intra-class variations can vary across classes depending upon their representative samples, affinity loss learns uniformed sized clusters. As visualized in Fig. 2, feature space within class disparities are observed for soft-max loss with minority classes occupying compact regions compared with their majority counterparts. In comparison, our proposed loss formulation is flexible, and allows learnt class prototypes to be equi-spaced and form uniformly shaped clusters. This reduces bias towards the less frequent cases and enhances the overall generalization capabilities, thus yielding a more discriminatively learnt feature space and an improved performance.

4.5. Robustness against Noisy Labels

For many real-world applications, the acquired data has noisy labels, and generalization of the learning methods against label noise is highly desirable [45, 53, 19]. To check the robustness of our proposed approach against noisy labels in the training data, we randomly flip the classes of MNIST training samples. The fraction of the miss-labelled samples is gradually increased from 10% to 50% with an increment of 10%. In order to avoid over-fitting on the noisy data, we deploy early stopping [60], and finish training when the performance on a held-out cross validation set starts to degrade. For comparison, we repeat all experiments using standard soft-max loss. The experimental results in Fig. 5 show that that the proposed Affinity loss performs better across the entire range of different noise levels. Although, the performance for both soft-max and affinity losses degrades with increasing noise factors, the proposed affinity loss shows more robustness, specially for larger noise ratios, with comparatively less performance degradation. The multi-centered learning in our loss provides flexibility to the noisy samples to associate themselves with class prototypes which are different from the non-noisy and
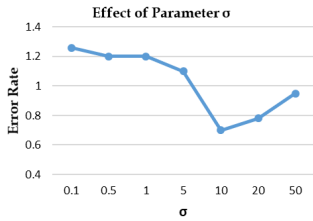
Figure 7. Effect of changing parameter σ that controls the spread of clusters. Results on imbalanced MNIST show that increasing the cluster variance above a certain point results in overlapped clusters and higher error rate.
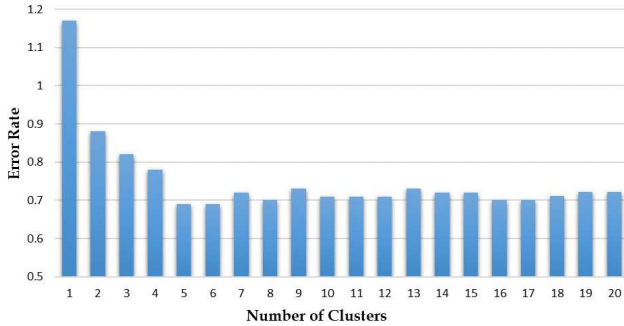
Distance d	Similarity	Performance
$\|a-b\|^1$	$\exp-(d/\sigma)$	99.3
$\|a-b\|^2$	$\exp-(d/\sigma)$	99.3
$\|a-b\|^1$	$1/(1+d)$	–
$\|a-b\|^2$	$1/(1+d)$	99.1

Table 4. Evaluation with different combinations of distance and similarity measures.

Figure 8. Performance for different number of clusters per class.

clean samples.

4.6. Ablation Study

Number of Cluster Centers: A unique aspect of the proposed affinity loss is its multi-centered learning which provides us the flexibility to have multiple class prototypes for each class. Here, we perform experiments on the imbalanced MNIST dataset (10% representation for first five digits), by gradually changing the number of representative prototypes m per class from 1 to 20. The experimental results in terms of error rates versus prototypes m in Fig. 8 show that the best performance is achieved for $m \geq 5$. Fewer prototypes per class ($m \leq 5$) yield relatively poor performance. The proposed method performs consistently the same when prototypes are increased beyond 5. Such multi-centered learning supports diversity in input samples. It is specifically helpful in scenarios with complex data distributions where large differences are observed amongst samples of the same class. Such diverse samples might not necessarily cluster around a single region, and could form multiple clusters by virtue of the proposed multi-centered learning mechanism. Furthermore, our experiments in Sec. 4.5 show that by providing flexible class prototypes, multi-centered learning proves an effective and robust scheme against noisy samples.

Cluster Spread σ: The parameter σ in Eq. 3 determines the cluster spread and helps achieve uniform intra-class variations. Our 2D visualization of the learnt features in Fig. 2 demonstrate that the clusters for each class are uniformly sized for both the majority and minority classes. This is in contrast to the traditional soft-max loss, where shrinked

feature space regions are observed for the minority classes. For our proposed loss formulation, the size of the cluster is directly related with the value of parameter σ, with larger σ indicating larger variance for a cluster. We perform experiment on imbalanced MNIST dataset for different values of the the parameter $\sigma = \{.1, .5, 1, 5, 10, 20, 50\}$. The results in Fig. 7 show that the optimal performance is achieved for values of σ between 5 and 20. Very high values of σ results in larger cluster spreads causing overlaps and confusion amongst classes and lower classification performance.

Distance and Similarity Metrics: Our original affinity loss formulation in Eq. 3 first computes the squared ℓ_2 distance between the feature $\mathbf{f}$ and class prototype $\mathbf{w}$, which is then converted to a similarity measure using the Gaussian metric. In this experiment, we evaluate different combinations of distance and similarity metrics. ℓ_1 and ℓ_2 metrics are used to compute distances, whereas Gaussian and inverse distance (defined by $\frac{1}{1+x}$) are the two similarity measures. We perform these experiments on imbalanced MNIST data (by retaining 10% samples for first five digits). Table 4 shows our evaluation results. The proposed scheme works well with all combinations except for ℓ_1 distance and Gaussian similarity, where it fails to converge. The best performance is achieved for Gaussian similarity in combination with squared ℓ_2 distance.

5. Conclusion

Class imbalance is ubiquitous in natural data and learning from such data is an unresolved challenge. The paper proposed a flexible loss formulation, aimed at producing a generalizable large margin classifier, to tackle class imbalance learning using deep networks. Based upon Euclidean space affinity defined with Bregmen divergence, the proposed loss jointly performs feature space clustering and max-margin classification. It enables learning uniformly sized equi-spaced clusters in the feature space, thus enhancing inter-class separability and reducing intra-class variations. The proposed scheme complements existing regularizers such as weight decay, and can be incorporated with different architectural backbones without incurring additional compute overhead. Experimental evaluations validate the effectiveness of the affinity loss for image classification benchmarks involving imbalanced data.

References

[1] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. Non-melanoma skin lesion classification using colour image data in a hierarchical k-nn classifier. In *Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on*, pages 358–361. IEEE, 2012. 6

[2] Lucia Ballerini, Robert B Fisher, Ben Aldridge, and Jonathan Rees. A color and texture based hierarchical k-nn approach to the classification of non-melanoma skin lesions. In *Color Medical Image Analysis*, pages 63–86. Springer, 2013. 5, 6

[3] Ricardo Barandela, E Rangel, José Salvador Sánchez, and Francesc J Ferri. Restricted decontamination for the imbalanced training sample problem. In *Iberoamerican Congress on Pattern Recognition*, pages 424–431. Springer, 2003. 2

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018. 5

[5] Cristiano L Castro and Antônio P Braga. Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. *IEEE transactions on neural networks and learning systems*, 24(6):888–899, 2013. 2

[6] Jia-Ren Chang and Yong-Sheng Chen. Batch-normalized maxout network in network. *arXiv preprint arXiv:1511.02583*, 2015. 6

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 2

[8] Binghui Chen, Weihong Deng, and Junping Du. Noisy softmax: Improving the generalization ability of dcnn via postponing the early softmax saturation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5372–5381, 2017. 7

[9] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 3

[10] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018. 1, 2, 4

[11] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou. Marginal loss for deep face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2006–2014. IEEE, 2017. 7

[12] Gamaleldin F Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. *arXiv preprint arXiv:1803.05598*, 2018. 3

[13] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *null*, pages 1735–1742. IEEE, 2006. 6

[14] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, pages 878–887. Springer, 2005. 2

[15] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 2

[16] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998. 2

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 7, 2017. 6

[18] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5375–5384, 2016. 2, 7

[19] Chen Huang, Chen Change Loy, and Xiaoou Tang. Discriminative sparse neighbor approximation for imbalanced learning. *IEEE transactions on neural networks and learning systems*, 29(5):1503–1513, 2018. 2, 7

[20] Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung. Classification of imbalanced data by combining the complementary neural network and smote algorithm. In *International Conference on Neural Information Processing*, pages 152–159. Springer, 2010. 2

[21] Taeho Jo and Nathalie Japkowicz. Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter*, 6(1):40–49, 2004. 2

[22] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 103–112, 2019. 2

[23] Salman Khan, Hossein Rahmani, Syed Afaq Ali Shah, and Mohammed Bennamoun. A guide to convolutional neural networks for computer vision. *Synthesis Lectures on Computer Vision*, 8(1):1–207, 2018. 2

[24] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Cost sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 2017. 2, 3, 6, 7

[25] Bartosz Krawczyk, Micha Woniak, and Gerald Schaefer. Cost-sensitive decision tree ensembles for effective imbalanced classification. *Applied Soft Computing*, 14:554 – 562, 2014. 2

[26] M KUBAT. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the 14th International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann, 1997. 2

[27] Steve Lawrence, Ian Burns, Andrew Back, Ah Chung Tsoi, and C Lee Giles. Neural network classification and prior class probabilities. In *Neural networks: Tricks of the trade*, pages 295–309. Springer, 2012. 2

[28] Gary B. Huang Erik Learned-Miller. Labeled faces in the wild: Updates and new reporting procedures. Technical Report UM-CS-2014-003, University of Massachusetts, Amherst, May 2014. 5

[29] Chen-Yu Lee, Patrick W Gallagher, and Zhuowen Tu. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 464–472, 2016. 6

[30] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. 2015. 6

[31] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul MB Vitányi. The similarity metric. *IEEE transactions on Information Theory*, 50(12):3250–3264, 2004. 4

[32] Jingtuo Liu, Yafeng Deng, Tao Bai, Zhengping Wei, and Chang Huang. Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*, 2015. 7

[33] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746. IEEE, 2017. 1, 2, 4, 7

[34] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning*, pages 507–516, 2016. 1, 2, 3, 7

[35] Yu Liu, Guanglu Song, Jing Shao, Xiao Jin, and Xiaogang Wang. Transductive centroid projection for semi-supervised large-scale recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–86, 2018. 3

[36] Iacopo Masi, Anh Tun Trn, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision*, pages 579–596. Springer, 2016. 7

[37] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of IEEE Intl Conf. on Computer Vision and Pattern Recognition (CVPR-W 2017)*, Honolulu, Hawaii, June 2017. 5, 6

[38] Aamir Mustafa, Salman Khan, Munawar Hayat, Roland Goecke, Jianging Shen, and Ling Shao. Adversarial defense by restricting the hidden space of deep neural networks. *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 2

[39] Wing WY Ng, Guangjun Zeng, Jiangjun Zhang, Daniel S Yeung, and Witold Pedrycz. Dual autoencoders features for imbalance classification problem. *Pattern Recognition*, 60:875–889, 2016. 2

[40] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *Proceedings of the British Machine Vision Conference*, pages 6–14, 2015. 7

[41] Michael D Richard and Richard P Lippmann. Neural network classifiers estimate bayesian a posteriori probabilities. *Neural computation*, 3(4):461–483, 1991. 2

[42] C.D. Castillo V.M. Patel R. Chellappa D.W. Jacobs S. Sengupta, J.C. Cheng. Frontal to profile face verification in the wild. In *IEEE Conference on Applications of Computer Vision*, February 2016. 5

[43] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2, 6, 7

[44] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pages 467–482. Springer, 2016. 2

[45] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622. ACM, 2008. 7

[46] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017. 2, 4

[47] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, pages 1988–1996, 2014. 2, 7

[48] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900, 2015. 7

[49] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014. 7

[50] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Web-scale training for face identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2746–2754, 2015. 7

[51] Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288, 2009. 2

[52] Kai Ming Ting. A comparative study of cost-sensitive boosting algorithms. In *In Proceedings of the 17th International Conference on Machine Learning*. Citeseer, 2000. 2

[53] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pages 3630–3638, 2016. 7

[54] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 7

[55] Shoujin Wang, Wei Liu, Jia Wu, Longbing Cao, Qinxue Meng, and Paul J Kennedy. Training deep neural networks on imbalanced data sets. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 4368–4374. IEEE, 2016. 2

[56] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017. 2

[57] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016. 2, 7

[58] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 529–534. IEEE, 2011. 5

[59] Yue Wu, Hongfu Liu, Jun Li, and Yun Fu. Deep face recognition with center invariant loss. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 408–414. ACM, 2017. 7

[60] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007. 7

[61] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for deep face recognition with long-tail data. *arXiv preprint arXiv:1803.09014*, 2018. 7

[62] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016. 6

[63] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418, 2017. 7

[64] Zhi-Hua Zhou and Xu-Ying Liu. On multi-class cost-sensitive learning. *Computational Intelligence*, 26(3):232–257, 2010. 2