Class Rectification Hard Mining for Imbalanced Deep Learning

Qi Dong
Queen Mary University of London
q.dong@qmul.ac.uk

Shaogang Gong
Queen Mary University of London
s.gong@qmul.ac.uk

Xiatian Zhu
Vision Semantics Ltd.
eddy@visionsemantics.com

Abstract

Recognising detailed facial or clothing attributes in images of people is a challenging task for computer vision, especially when the training data are both in very large scale and extremely imbalanced among different attribute classes. To address this problem, we formulate a novel scheme for batch incremental hard sample mining of minority attribute classes from imbalanced large scale training data. We develop an end-to-end deep learning framework capable of avoiding the dominant effect of majority classes by discovering sparsely sampled boundaries of minority classes. This is made possible by introducing a Class Rectification Loss (CRL) regularising algorithm. We demonstrate the advantages and scalability of CRL over existing state-of-the-art attribute recognition and imbalanced data learning models on two large scale imbalanced benchmark datasets, the CelebA facial attribute dataset and the X-Domain clothing attribute dataset.

1. Introduction

Automatic recognition of person attributes in images, e.g. clothing category and facial characteristics, is very useful [17, 15], but also challenging due to: (1) Very large scale training data with significantly imbalanced distributions on annotated attribute data [1, 6, 21], with clothing and face attributes typically exhibiting a power-law distribution (Figure 1). This makes model learning biased towards well-labelled attribute classes (the *majority classes*) resulting in poor performance against sparsely-labelled classes (the *minority classes*) [20], known as the *imbalanced class learning* problem [20]. (2) Subtle discrepancy between different fine-grained attributes, e.g. "Woollen-Coat" can appear very similar to "Cotton-Coat", whilst "Mustache" may be visually indistinct (Figure 1). To recognise such subtle attribute differences, model training *assumes* a large collection of *balanced* training image data [7, 50].

There have been studies on how to solve the general imbalanced data learning problem including re-sampling [4, 33, 36] and cost-sensitive weighting [44, 43]. However, these methods can suffer from either *over-sampling* which leads to model overfitting and/or introducing noise, or *down-sampling* which loses valuable data. These classical imbalanced learning models rely typically on hand-crafted features, without deep learning's capacity for exploiting a very large pool of imagery data from diverse sources to learn more expressive representations [41, 39, 27, 3]. However, deep learning is likely to suffer even more from imbalanced data distribution [51, 24, 25, 21] and deep learning of imbalanced data is currently under-studied. This is partly due to that popular image datasets for deep learning, e.g. ILSVRC, do not exhibit significant class imbalance due to careful data filtering and selection during the construction process (Table 1). The problem becomes very challenging for deep learning of clothing or facial attributes (Figure 1). In particular, when a large scale training data are drawn from online Internet sources [7, 22, 31, 32], image attribute distributions are likely to be extremely imbalanced (see Table 1). For example, the data sampling size ratio between the minority and majority classes (imbalance ratio) in the X-Domain clothing attribute dataset [7] is 1:4,162, with the smallest minority and largest majority class having 24 and 99, 885 images respectively.

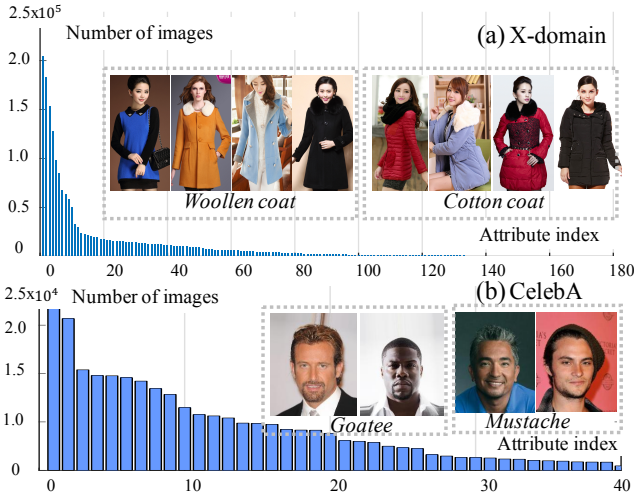

Figure 1. Imbalanced training data distribution: (a) clothing attributes (X-Domain [7]), (b) facial attributes (CelebA [32]).

Table 1. Comparing large scale datasets in terms of training data imbalance. Metric: the size ratio of smallest and largest classes. These numbers are based on the standard train data split if available, otherwise on the whole dataset. For COCO [29], no specific numbers are available for calculating between-class imbalance ratios, mainly because the COCO images often contain simultaneously multiple classes of objects and also multiple instances of a specific class.

Datasets	ILSVRC2012-14 [37]	COCO [29]	VOC2012 [12]	CIFAR-100 [26]	Caltech 256 [18]	CelebA [32]	DeepFashion [31]	X-Domain [7]
Imbalance ratio	1 : 2	-	1 : 13	1 : 1	1 : 1	1 : 43	1 : 733	**1 : 4162**

This work addresses the problem of deep learning on large scale imbalanced person attribute data for multi-label attribute recognition. Other deep models for imbalanced data learning exist [51, 24, 36, 25]. These models shall be considered as end-to-end deep feature learning and classifier learning. For over-sampling and down-sampling, a special training data re-sampling pre-process may be needed prior to deep model learning. They are ineffective for deep learning of imbalanced data (see evaluations in Sec. 3). More recently, a Large Margin Local Embedding (LMLE) method [21] was proposed to enforce the local cluster structure of per class distribution in the deep learning process so that minority classes can better maintain their own structures in the feature space. The LMLE has a number of fundamental drawbacks including disjoint feature and classification optimisation, offline clustering of training data *a priori* to model learning, and quintuplet construction updates.

This work presents a novel *end-to-end* deep learning approach to modelling multi-label person attributes, clothing or facial, given a large scale webly-collected image data pool with significantly imbalanced attribute data distributions. The **contributions** of this work are: (1) We propose a novel model for deep learning of very large scale imbalanced data based on *batch-wise incremental hard mining* of hard-positives and hard-negatives from minority attribute classes alone. This is in contrast to existing attribute recognition methods [7, 22, 31, 10, 50] which either assume balanced training data or simply ignore the problem. Our model performs an end-to-end feature representation and multi-label attribute classification joint learning. (2) We formulate a *Class Rectification Loss* (CRL) regularising algorithm. This is designed to explore the per batch sampled hard-positives and hard-negatives for improving minority class learning with batch-balance updated deep features. Crucially, this loss rectification is correlated explicitly with batch-wise (small data pool) iterative model optimisation therefore achieving incremental imbalanced data learning for all attribute classes. This is in contrast to LMLE's global clustering of the entire training data (large data pool) and ad-hoc estimation of cluster size. Moreover, given our batch-balancing hard-mining approach, the proposed CRL is independent to the overall training data size, therefore very scalable to large scale training data. Our extensive experiments on two large scale datasets CelebA [32] and X-Domain [7] against 11 different models including 7 state-of-the-art deep attribute models demonstrate the advantages of the proposed method.

Related Work. *Imbalanced Data Learning.* There are two classic approaches to learning from imbalanced data, (1) *Class re-sampling*: Either down-sampling the majority class or over-sampling the minority class or both [4, 11, 19, 20, 33, 36]. However, over-sampling can easily introduce undesirable noise and also risk from overfitting. Down-sampling is thus often preferred, but this may suffer from losing valuable information [11]. (2) *Cost-sensitive learning*: Assigning higher misclassification costs to the minority classes as compared to the majority classes [44, 49, 5, 51, 43], or regularising the cross-entropy loss to cope with the imbalanced positive and negative class distribution [40]. For this kind of data biased learning, most commonly adopted in deep models is positive data augmentation, e.g. to learn a deep representation embedding the local feature structures of minority labels [21]. *Hard Mining.* Negative mining has been used for pedestrian detection [14], face recognition [38], image categorisation [35, 46, 9], unsupervised visual representation learning [48]. Instead of general negative mining, the rational for mining *hard* negatives (unexpected) is that they are more informative than *easy* negatives (expected). Hard negative mining enables the model to improve itself quicker and more effectively with less data. Similarly, model learning can also benefit from mining hard positives (unexpected). In our model learning we *only* consider hard mining on the minority classes for efficiency therefore our batch-balancing hard mining strategy differs significantly from that of LMLE [21] in that: (1) The LMLE requires to exhaustively search the entire training set and thus less scalable to large sized data due to computational cost; (2) Hard mining in LMLE is on *all* classes, both the minority and the majority classes, therefore not strictly focused on imbalanced learning of the minority classes thus more expensive whilst less effective. *Deep Learning of Person Attributes.* Personal clothing and/or facial attributes are key to person description. Deep learning have been exploited for clothing [7, 22, 31, 10, 47] and facial attribute recognition [32, 50] due to the availability of large scale datasets and deep models' capacity for learning from large sized data. However, these methods mostly ignore the significantly imbalanced class data distributions, resulting in suboptimal model learning for the minority classes. One exception is the LMLE model [21] which explicitly considers the imbalanced attribute class learning challenge. In contrast to our end-to-end deep learning model in this work, LMLE is not end-to-end learning and suffers from poor scalability and suboptimal optimisation. This is due to LMLE's need for very expensive quintuplet construction and pre-clustering (suboptimal) on the entire training data, resulting in separated feature and classifier learning.

2. Class Rectification Deep Learning

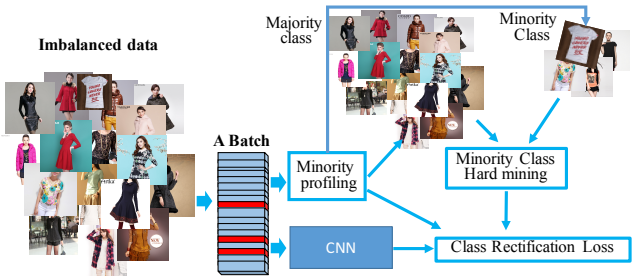

Figure 2. Overview of our Class Rectification Loss (CRL) regularising approach for deep end-to-end imbalanced data learning.

We wish to construct a deep model capable of recognising multi-labelled person attributes $\{z_j\}_{j=1}^{n_{\text{attr}}}$ in images, with a total of n_{attr} different attribute categories, each category z_j having its respective value range Z_j, e.g. multi-valued (1-in-N) clothing category or binary-valued (1-in-2) facial attribute. Suppose that we have a collection of n training images $\{\boldsymbol{I}_i\}_{i=1}^{n}$ along with their attribute annotation vectors $\{\boldsymbol{a}_i\}_{i=1}^{n}$, and $\boldsymbol{a}_i = [a_{i,1}, \ldots, a_{i,j}, \ldots, a_{i,n_{\text{attr}}}]$ where $a_{i,j}$ refers to the j-th attribute value of the image $\boldsymbol{I}_i$. The number of image samples available for different attribute classes varies greatly (Figure 1) therefore poses a significant *imbalanced data* distribution challenge to model learning. Most attributes are *localised* to image regions, even though the location information is not provided in the annotation (*weakly labelled*). Intrinsically, this is a *multi-label* recognition problem since the n_{attr} attributes may co-exist in every person image. To that end, we consider to jointly learn *end-to-end* features and *all* the attribute classifiers given imbalanced image data. Our method can be readily incorporated with the classification loss function (e.g. Cross-entropy loss) of standard CNNs without the need for a new optimisation algorithm (Fig. 2).

Cross-entropy Classification Loss. For multi-class classification CNN model training (CNN model details in "Network Architecture", Sec. 3.1 and 3.2), one typically considers the Cross-entropy loss function by firstly predicting the j-th attribute posterior probability of image $\boldsymbol{I}_i$ over the ground truth $a_{i,j}$:

$$p(y_{i,j} = a_{i,j}|\boldsymbol{x}_{i,j}) = \frac{\exp(\boldsymbol{W}_j^\top \boldsymbol{x}_{i,j})}{\sum_{k=1}^{|Z_j|} \exp(\boldsymbol{W}_k^\top \boldsymbol{x}_{i,j})} \quad (1)$$

where $\boldsymbol{x}_{i,j}$ refers to the feature vector of $\boldsymbol{I}_i$ for j-th attribute, and $\boldsymbol{W}_k$ is the corresponding prediction function parameter. We then compute the overall loss on a batch of n_{bs} images as the average additive summation of attribute-level loss with equal weight:

$$l_{\text{ce}} = -\frac{1}{n_{\text{bs}}} \sum_{i=1}^{n_{\text{bs}}} \sum_{j=1}^{n_{\text{attr}}} \log\Big(p(y_{i,j} = a_{i,j}|\boldsymbol{x}_{i,j})\Big) \quad (2)$$

However, given highly imbalanced image samples on different attribute classes, model learning by the conventional classification loss is suboptimal. To address this problem, we reformulate the model learning objective loss function by mining explicitly in each batch of training data both hard positive and hard negative samples for every minority attribute class. Our objective is to rectify incrementally per batch the class bias in model learning so that the features are less biased towards the over-sampled majority classes and more sensitive to the class boundaries of under-sampled minority classes.

2.1. Minority Class Hard Mining

We wish to impose minority-class hard-samples as constraints on the model learning objective. Different from the approach adopted by the LMLE model [21] which aims to preserve the local structures of *both* majority and minority classes by global sampling of the entire training dataset, we explore *batch-based* hard-positive and hard-negative mining for the minority classes *only*. We do not assume the local structures of minority classes can be estimated from global clustering before model learning. To that end, we consider the following steps for handling data imbalance.

Batch Profiling of Minority and Majority Classes. In each training batch, we profile to discover the minority and majority classes. Given a batch of n_{bs} training samples, we profile the attribute class distribution $\boldsymbol{h}^j = [h_1^j, \ldots, h_k^j, \ldots h_{|Z_i|}^j]$ over Z_j for each attribute j, where h_k^j denotes the number of training samples with the j-th attribute class value assigned to k. Then, we sort h_k^j in the descent order. As such, we define minority classes *in this batch* as those classes C_{min}^i with the smallest number of training samples, with the condition that

$$\sum_{k \in C_{\text{min}}^j} h_k^j < 0.5 n_{\text{bs}}. \quad (3)$$

That is, all minority classes only contribute to less than half of the total data samples in this batch. The remaining classes are deemed as the majority classes.

We then exploit a minority class hard mining scheme to facilitate additional loss constraints in model learning[1]. To that end, we consider two approaches: (I) Minority class-level hard mining (Fig. 3(left)), (II) minority instance-level hard mining (Fig. 3(right)).

(I) Minority Class-Level Hard Samples. At the class level, for a specific minority class c of attribute j, we refer "hard-positives" to those images $\boldsymbol{x}_{i,j}$ from class c ($a_{i,j} = c$ with $a_{i,j}$ denoting the attribute j ground truth label of $\boldsymbol{x}_{i,j}$) given *low* discriminative scores $p(y_{i,j} = c|\boldsymbol{x}_{i,j})$ on class c

[1] We consider only those minority classes having at least two sample images in each batch, ignoring those minority classes having only one sample image or none. This enables triplet loss based learning.

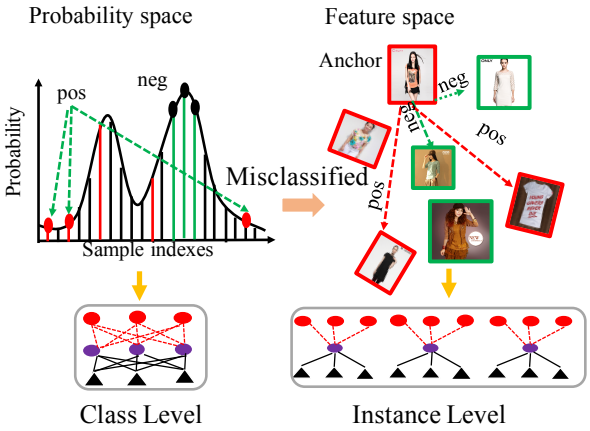

Figure 3. Illustration of the proposed minority class hard mining.

by the model, i.e. *poor* recognitions. Conversely, by "hard-negatives", we refer to those images $\boldsymbol{x}_{i,j}$ from other classes ($a_{i,j} \neq c$) given *high* discriminative scores on class c by the model, i.e. *obvious* mistakes. Formally, we define them as:

$$\mathcal{P}^{\text{cls}}_{c,j} = \{\boldsymbol{x}_{i,j} | a_{i,j} = c,\ \text{low}\ p(y_{i,j} = c|\boldsymbol{x}_{i,j})\} \quad (4)$$

$$\mathcal{N}^{\text{cls}}_{c,j} = \{\boldsymbol{x}_{i,j} | a_{i,j} \neq c,\ \text{high}\ p(y_{i,j} = c|\boldsymbol{x}_{i,j})\} \quad (5)$$

where $\mathcal{P}^{\text{cls}}_{c,j}$ and $\mathcal{N}^{\text{cls}}_{c,j}$ denote the hard positive and negative sample sets of a minority class c of attribute j.

(II) Minority Instance-Level Hard Samples. At the instance level, we consider hard positives and negatives for each specific sample instance $\boldsymbol{x}_{i,j}$ from a minority class c of attribute j, i.e. with $a_{i,j} = c$. We define "hard-positives" of $\boldsymbol{x}_{i,j}$ as those class c images $\boldsymbol{x}_{k,j}$ ($a_{k,j} = c$) misclassified ($\hat{a}_{k,j} \neq c$ with $\hat{a}_{k,j}$ denoting the attribute j predicted label of $\boldsymbol{x}_{k,j}$) by the current model with *large* distances (low matching scores) from $\boldsymbol{x}_{i,j}$ in the feature space. "Hard-negatives" are those images $\boldsymbol{x}_{k,j}$ not from class c ($a_{k,j} \neq c$) with *small* distances (high matching scores) to $\boldsymbol{x}_{i,j}$ in the feature space. We define them as:

$$\mathcal{P}^{\text{ins}}_{i,c,j} = \{\boldsymbol{x}_{k,j} | a_{k,j} = c,\ \hat{a}_{k,j} \neq c,\ \text{large dist}(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{k,j})\} \quad (6)$$

$$\mathcal{N}^{\text{ins}}_{i,c,j} = \{\boldsymbol{x}_{k,j} | a_{k,j} \neq c,\ \text{small dist}(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{k,j})\} \quad (7)$$

where $\mathcal{P}^{\text{ins}}_{i,c,j}$ and $\mathcal{N}^{\text{ins}}_{i,c,j}$ are the hard positive and negative sample sets of a minority class c instance $\boldsymbol{x}_{i,j}$ in attribute j, and dist$(\cdot)$ is the L_2 distance metric.

Hard Mining. Intuitively, mining hard-positives enables the model to discover and expand sparsely sampled minority class boundaries, whilst mining hard-negatives aims to improve the margins of minority class boundaries corrupted by visually very similar imposter classes, e.g. significantly overlapped outliers. To facilitate and simplify model training, we adopt the following mining strategy. At training time, for a minority class c of attribute j (or a minority class instance $\boldsymbol{x}_{i,j}$) in each training batch data, we select K hard-positives as the bottom-K scored on c (or bottom-K (largest) distances to $\boldsymbol{x}_{i,j}$), and K hard-negatives as the top-K scored on c (or top-K (smallest) distance to $\boldsymbol{x}_{i,j}$), given the current feature space and classification model. This hard mining strategy allows our model optimisation to concentrate particularly on either poor recognitions or obvious mistakes. This not only reduces the model optimisation complexity by soliciting fewer learning constraints, but also minimises computing cost. It may seem that some discriminative information is lost by doing so. However, it should be noted that we perform hard-mining *independently* in each batch and *incrementally* over successive batches. Therefore, such seemingly-ignored information are considered over the learning iterations. Importantly, this proposed batch-wise hard-mining avoids the global sampling on the entire training data as required by LMLE [21] which can suffer from both negative model learning due to inconsistency between up-to-date deep features and out-of-date cluster boundary structures, and high computational cost in quintuplet updating. In contrast, our model can be learned directly by conventional batch-based classification optimisation algorithms using stochastic gradient descent, with no need for complex modification required by the quintuplet based loss in the LMLE model [21].

2.2. Class Rectification Loss

In deep feature representation model learning, the key is to discover latent boundaries for individual classes and surrounding margins between different classes in the feature space. To this end, we introduce a Class Rectification Loss (CRL) regularisation l_{crl} to rectify the learning bias from the conventional Cross-entropy classification loss function (Eqn. (2)) given class-imbalanced attribute data:

$$l_{\text{bln}} = l_{\text{crl}} + l_{\text{ce}} \quad (8)$$

where l_{crl} is computed from the hard positive and negative samples of the minority classes. We further explore three different options to formulate l_{crl}.

(I) Class Rectification by Relative Comparison. Firstly, we exploit the general learning-to-rank idea [30], and in particular the triplet based loss. Considering the small number of training samples in minority classes, it is sensible to make full use of them in order to effectively handle the underlying model learning bias. Therefore, we regard each image of these minority classes as an "anchor" to quantitatively compute the batch balancing loss regularisation. Specifically, for each anchor ($\boldsymbol{x}_{a,j}$), we first construct a set of triplets based on the mined top-K hard-positives and hard-negatives associated with the corresponding attribute class c of attribute j, i.e. class-level hard mimng, or the sample instance itself $\boldsymbol{x}_{a,j}$, i.e. instance-level hard mining. In this way, we form at most K^2 triplets $T = \{(\boldsymbol{x}_{a,j}, \boldsymbol{x}_{p,j}, \boldsymbol{x}_{n,j})_k\}_{k=1}^{K^2}$ w.r.t. $\boldsymbol{x}_{a,j}$, and a total of at

most $|X_{\text{min}}| \times n_{\text{attr}} \times K^2$ triplets T for all the anchors X^i_{min} across all the minority classes. We formulate the following triplet ranking loss function to impose a class balancing constraint in model learning:

$$l_{\text{crl}} = \frac{\sum_T \max\left(0,\ m_j + \text{dist}(\boldsymbol{x}_{a,j}, \boldsymbol{x}_{p,j}) - \text{dist}(\boldsymbol{x}_{a,j}, \boldsymbol{x}_{n,j})\right)}{|T|} \tag{9}$$

where m_j denotes the class margin of attribute j in feature space, $\text{dist}(\cdot)$ is the L_2 distance. We set the class margin for each attribute i as

$$m_j = \frac{2\pi}{|Z_j|} \tag{10}$$

with $|Z_j|$ the number of all possible values for attribute j.

(II) Class Rectification by Absolute Comparison. Secondly, we consider to enforce absolute distance constraints on positive and negative pairs of the minority classes, inspired by the contrastive loss [8]. Specifically, for each instance $\boldsymbol{x}_{i,j}$ in a minority class c of attribute j, we use the mined hard sets to build positive $P^+ = \{\boldsymbol{x}_{i,j}, \boldsymbol{x}_{p,j}\}$ and negative $P^- = \{\boldsymbol{x}_{i,j}, \boldsymbol{x}_{n,j}\}$ pairs in each training batch. Intuitively, we require the positive pairs to be at close distances whist the negative pairs to be far away. Thus, we define the CRL regularisation as

$$\begin{aligned} l_{\text{crl}} = 0.5 * \Big(& \frac{\sum_{P^+} \text{dist}(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{p,j})^2}{|P^+|} + \\ & \frac{\sum_{P^-} \max\left(m_{\text{apc}} - \text{dist}(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{n,j}), 0\right)^2}{|P^-|} \Big) \end{aligned} \tag{11}$$

where m_{apc} controls the between-class margin ($m_{\text{apc}} = 1$ in our experiments). This constraint aims to optimise the boundary of the minority classes by incremental separation from the overlapping (confusing) majority class instances by per batch iterative optimisation.

(III) Class Rectification by Distribution Comparison. Thirdly, we formulate class rectification on the minority class instances by modelling the *distribution* of positive and negative pairs constructed as in the case of "Absolute Comparison" described above. In the spirit of [45], we represent the distribution of positive P^+ and negative P^- pair sets with histograms $H^+ = [h^+_1, \cdots, h^+_\tau]$ and $H^- = [h^-_1, \cdots, h^-_\tau]$ of τ uniformly spaced bins $[b_1, \cdots, b_\tau]$. We compute the positive histogram H^+ as

$$h^+_t = \frac{1}{|P^+|} \sum_{(i,j) \in P^+} \varsigma_{i,j,t} \tag{12}$$

where

$$\varsigma_{i,j,t} = \begin{cases} \frac{\text{dist}(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{p,j}) - b_{t-1}}{\Delta}, & \text{if } \text{dist}(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{p,j}) \in [b_{t-1}, b_t] \\ \frac{b_{t+1} - \text{dist}(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{p,j})}{\Delta}, & \text{if } \text{dist}(\boldsymbol{x}_{i,j}, \boldsymbol{x}_{p,j}) \in [b_t, b_{t+1}] \\ 0. & \text{otherwise} \end{cases} \tag{13}$$

and Δ defines the step between two adjacent bins. Similarly, the negative histogram H^- can also be computed. To enable the minority classes distinguishable from the overwhelming majority classes, we enforce the two histogram distributions as disjoint as possible. We then define the CRL regularisation loss by how much overlapping between these two histogram distributions:

$$l_{\text{crl}} = \sum_{t=1}^{\tau} \left(h^+_t \sum_{k=1}^{t} h^-_k\right) \tag{14}$$

Statistically, this CRL histogram loss measures the probability that the distance of a random negative pair is smaller than that of a random positive pair. This distribution based CRL aims to optimise a model towards mining the minority class boundary areas in a non-deterministic manner. In our evaluation (Sec. 3.3), we compared the effect of these three different CRL considerations. By default, we deploy the Relative Comparison formulation in our experiments.

Remarks. Due to the batch-wise design, the balancing effect by our proposed regularisor is propagated through the whole training time in an incremental manner. The CRL approach shares a similar principle to Batch Normalisation [23] for easing network optimisation. In hard mining, we do not consider anchor points from the majority classes as in the case of LMLE [21]. Instead, our method employs a classification loss to learn features for discriminating the majority classes based on that the majority classes are well-sampled for learning class discrimination. Focusing the CRL *only* on the minority classes makes our model computationally more efficient. Moreover, the computational complexity for constructing quintuplets for LMLE and updating class clustering globally is $n_{\text{attr}} \times (k \times O(n) \times 2^{\Omega(\sqrt{n})}) + O(n^2)$ where Ω is the lower bound complexity and O the upper bound complexity, that is, super-polynomially proportionate to the overall training data size n, e.g. over $150,000$ in our attribute recognition problem. In contrast, CRL loss is linear to the batch size, typically in 10^2, independent to the overall training size (also see "Model Training Time" in the experiments).

3. Experiments

Datasets & Performance Metric. As shown in Table 1, both CelebA and X-Domain datasets are highly imbalanced. For that reason, we selected these two datasets for our evaluations. The CelebA [32] facial attribute dataset has 202,599 web images from 10,177 person identities with per person on average 20 images. Each face image is annotated by 40 binary attributes. The X-Domain [7] clothing attribute dataset[2] consists of 245,467 shop images from online re-

[2]We did not select the DeepFashion [31] dataset for our evaluation because this dataset is relatively well balanced compared to X-Domain (Table 1), due to the strict data cleaning process applied.

Table 2. Facial attributes recognition on the CelebA dataset [32]. *: Imbalanced data learning models. Metric: Class-balanced accuracy, i.e. mean sensitivity (%). CRL(C/I): CRL with Class/Instance level hard mining. The 1st/2nd best results are highlighted in red/blue.

Methods \ Attributes	Attractive	Mouth Open	Smiling	Wear Lipstick	High Cheekbones	Male	Heavy Makeup	Wavy Hair	Oval Face	Pointy Nose	Arched Eyebrows	Black Hair	Big Lips	Big Nose	Young	Straight Hair	Brown Hair	Bags Under Eyes	Wear Earrings	No Beard	
Imbalance ratio (1:x)	1	1	1	1	1	1	2	2	3	3	3	3	3	3	4	4	4	4	4	5	
Triplet-kNN [38]	83	92	92	91	86	91	88	77	61	61	73	82	55	68	75	63	76	63	69	82	
PANDA [50]	85	93	**98**	**97**	89	**99**	**95**	78	66	67	77	84	56	72	78	66	85	67	77	87	
ANet [32]	**87**	**96**	97	95	**89**	**99**	96	**81**	**67**	69	76	**90**	57	**78**	84	69	**83**	70	**83**	93	
DeepID2 [42]	78	89	89	92	84	94	88	73	63	66	77	83	**62**	73	76	65	79	74	75	88	
Over-Sampling* [24]	77	89	90	92	84	95	87	70	63	67	**79**	84	61	73	75	66	82	73	76	88	
Down-Sampling* [34]	78	87	90	91	80	90	89	70	58	63	70	80	61	76	80	61	76	71	70	88	
Cost-Sensitive* [20]	78	89	90	91	85	93	89	75	64	65	78	85	61	74	75	67	84	74	76	88	
LMLE* [21]	**88**	**96**	**99**	**99**	**92**	**99**	**98**	**83**	**68**	72	**79**	**92**	60	**80**	**87**	**73**	**87**	73	**83**	**96**	
CRL(C)*	80	92	90	93	85	**96**	88	**81**	**68**	**77**	**80**	88	**68**	77	**85**	**76**	82	**79**	**82**	91	
CRL(I)*	83	**95**	93	94	89	**96**	84	79	66	**73**	**80**	**90**	**68**	**80**	84	73	**86**	**80**	**83**	**94**	

Methods \ Attributes	Bangs	Blond Hair	Bushy Eyebrows	Wear Necklace	Narrow Eyes	5 oclock Shadow	Receding Hairline	Wear Necktie	Eyeglasses	Rosy Cheeks	Goatee	Chubby	Sideburns	Blurry	Wear Hat	Double Chin	Pale Skin	Gray Hair	Mustache	Bald	**Average**
Imbalance ratio (1:x)	6	6	6	7	8	8	11	13	14	14	15	16	17	18	19	20	22	23	24	43	
Triplet-kNN [38]	81	81	68	50	47	66	60	73	82	64	73	64	71	43	84	60	63	72	57	75	**72**
PANDA [50]	92	91	74	51	51	76	67	85	88	68	84	65	81	50	90	64	69	79	63	74	**77**
ANet [32]	90	90	**82**	59	57	81	70	79	95	76	86	70	79	56	90	68	77	85	61	73	**80**
DeepID2 [42]	91	90	78	70	64	85	81	83	92	86	90	**81**	89	74	90	83	81	90	88	93	**81**
Over-Sampling* [24]	90	90	80	71	65	85	82	79	91	**90**	89	83	90	76	89	84	82	90	90	92	**82**
Down-Sampling* [34]	88	85	75	66	61	82	79	80	85	82	85	78	80	68	90	80	78	88	60	79	**78**
Cost-Sensitive* [20]	90	89	79	71	65	84	81	82	91	**92**	86	82	90	76	90	84	80	90	88	93	**82**
LMLE* [21]	**98**	**99**	**82**	59	59	82	76	**90**	**98**	78	**95**	79	88	59	**99**	74	80	91	73	90	**84**
CRL(C)*	93	91	**82**	**76**	**70**	**89**	**84**	84	**97**	87	92	83	**91**	**81**	94	**85**	**88**	**93**	**90**	**95**	**85**
CRL(I)*	**95**	**95**	**84**	**74**	**72**	**90**	**87**	**88**	96	88	**96**	**87**	**92**	**85**	**98**	**89**	**92**	**95**	**94**	**97**	**86**

tailers like *Tmall.com*. Each clothing image is annotated by ≤ 9 attribute categories and each category has a different set of values (mutually exclusive within each set) ranging from 6 (slv-len) to 55 (colour). In total, there are 178 distinctive attribute values in 9 categories (labels). For each attribute label, we adopted the *class-imbalanced* accuracy (i.e. mean sensitivity) as the model performance metric given imbalanced data [16, 21]. This additionally considers the class distribution statistics in performance measurement.

3.1. Evaluation on Imbalanced Face Attributes

Competitors. We compared CRL against 8 existing methods including 4 state-of-the-art deep models for facial attribute recognition on CelebA: (1) Over-Sampling [11], (2) Down-Sampling [11], (3) Cost-Sensitive [20], (4) Large Margin Local Embedding (LMLE) [21], (5) PANDA [50], (6) ANet [32], (7) Triplet-kNN [38], and (8) DeepID2 [42].

Training/Test Data Partition. We adopted the same data partition on CelebA as in [32, 21]: The first 162,770 images are used for training (10,000 images for validation), the following 19,867 images for training the SVM classifiers required by PANDA [50] and ANet [32] models, and the remaining 19,962 images for testing. Note that identities of all face images are non-overlapped in this partition.

Network Architecture & Parameter Settings. We adopted the five layers CNN network architecture of DeepID2 [42] as the basis for training all six imbalanced data learning methods including both our CRL models (C&I), the same for LMLE as reported in [21]. In addition to the DeepID2's shared FC_1 layer, for explicitly modelling the attribute specificness, in our CRL model we added a respective 64-dimensional FC_2 layer for each face attribute, in the spirit of multi-task learning [13, 2]. We set the learning rate at 0.001 to train our model from scratch on the CelebA face images. We fixed the decay to 0.0005 and the momentum to 0.9. Our CRL model converges after 200 epochs training with a batchsize of 128 images.

Comparative Evaluation. Facial attribute recognition performance comparisons are shown in Table 2. It is evident that CRL outperforms on average accuracy all competitors including the state-of-the-art attribute recognition models and imbalanced data learning methods Compared to the best non-imbalanced learning model DeepID2, CRL(I) improves average accuracy by 5%. Compared to the state-of-the-art imbalanced learning model LMLE, CRL(I) is better by 2% in average accuracy. Other classical imbalanced learning methods perform similarly to DeepID2. The performance drop by Down-Sampling is due to discarding useful data for balancing distributions. This demonstrates the importance of explicit imbalanced data learning, and the superiority of the proposed batch incremental class rectification hard mining approach to handling imbalanced data over

alternative methods. Figure 4 shows qualitative examples.

Figure 4. Examples (3 pairs) of facial attribute recognition (imbalance ratio in bracket). In each pair, DeepID2 missed both, whilst CRL identified the left image but failed the right image.

Model Performance vs. Data Imbalance Ratio. Figure 5 further shows the accuracy gain of six imbalanced learning methods. It can be seen that LMLE copes better with less imbalanced attributes (towards the left side in Figure 5), but degrades notably given higher data imbalance ratio. Also, LMLE performs worse than DeepID2 on more attributes towards the right of "Wear Necklace" in Figure 5, i.e. imbalance ratio greater than 1:7 in Table 2. In contrast, CRLs with both class-level (CRL(C)) and instance-level (CRL(I)) hard mining perform particularly well on attributes with high imbalance ratios. More importantly, even though CRL(I) only outperforms LMLE by 2% in average accuracy over all 40 attributes, this margin increases to 7% in average accuracy over the *20 most* imbalanced attributes. Moreover, on some of the very imbalanced attributes, CRL(I) outperforms LMLE by 21% on "Mustache" and 26% on "Blurry". Interestingly, the "Blurry" attribute is challenging due to its global characteristics therefore not defined by local features and very subtle, similar to the "Mustache" attribute (see Figure 4). This demonstrates that CRL is significantly better than LMLE in coping with severely imbalanced data learning. This becomes more evident with the X-domain clothing attributes (Sec. 3.2), mainly because given severe imbalanced data, it is difficult for LMLE to cluster effectively due to very few minority class samples, which leads to inaccurate classification feature learning.

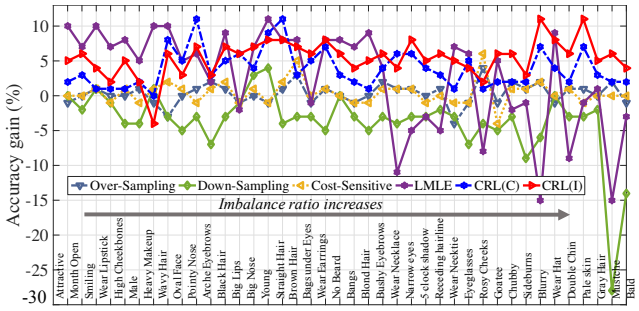

Figure 5. Performance gain over DeepID2 [42] by the six imbalanced learning methods on the 40 CelebA facial attributes [32]. Attributes sorted from left to right in increasing imbalance ratio.

Model Training Time. We also tested the training time cost of LMLE independently on an identical hardware setup as for CRL: LMLE took 388 hours to train whilst CRL (C/I) took 27/35 hours respectively with 11 times training costs advantage over LMLE in practice. Specifically, LMLE needs 4 rounds of quintuplets construction with each taking 96 hours, and 4 rounds of deep model learning with each taking 1 hour. In total, 4 * (96+1) = 388 hours.

3.2. Evaluation on Imbalanced Clothing Attributes

Competitors. In addition to the four imbalanced learning methods (Over-Sampling, Down-Sampling, Cost-Sensitive, LMLE[4]) used for face attribute evaluation, we also compared against four other state-of-the-arts clothing attribute recognition models: (1) Deep Domain Adaptation Network (DDAN) [7], (2) Dual Attribute-aware Ranking Network (DARN) [22], (3) FashionNet [31], and (4) Multi-Task Curriculum Transfer (MTCT) [10].

Training/Test Data Partition. We adopted the same data partition as in [22, 10]: Randomly selecting 165,467 clothing images for training and the remaining 80,000 for testing.

Network Architecture. We used the same network structure as the MTCT [10]. Specifically, this network is composited of five stacked NIN conv units [28] and n_{attr} parallel branches with each a three FC layers sub-network for modelling one of the n_{attr} attributes respectively, in the spirit of multi-task learning [13, 2].

Parameter Settings. We pre-trained a base model on ImageNet-1K at the learning rate 0.001, and then finetuned the CRL model on the X-Domain clothing images at the same rate 0.001. We fixed the decay to 0.0005 and the momentum to 0.9. The CRL model converges after 150 epochs. The batchsize is 256.

Comparative Evaluation. Table 3 shows the comparative evaluation of 10 different models on the X-Domain benchmark dataset. It is evident that CRL(I) surpasses all other models on all attribute categories. This shows the significant superiority and scalability of the class rectification hard mining with batch incremental approach in coping with extremely imbalanced attribute data, with the maximal imbalance ratio $4,162$ vs. 43 in CelebA attributes (Figure 1). A lack of explicit imbalanced learning mechanism in other models such as DDAN, FashionNet, DARN and MTCT suffers notably. Among the 6 models designed for imbalance data learning, we can observe similar trends as in face attribute recognition on CelebA. Whilst LMLE improves notably on classic imbalanced data learning methods, it remains inferior to CRL(I) by significant margins (4% in accuracy over all attributes).

Model Effectiveness in Mitigating Data Imbalance. We compared the relative performance gain of the 6 different imbalanced data learning models (Down-Sampling was excluded due to poor performance) against the MTCT (as the

[4]We trained an independent LMLE CNN model for each attribute label. This is because the quintuplets construction over all attribute labels is prohibitively expensive in terms of computing cost.

Table 3. Clothing attributes recognition on the X-Domain dataset. * Imbalanced data learning models. Metric: Class-balanced accuracy, i.e. mean sensitivity (%).. CRL(C/I): CRL with Class/Instance level hard mining. Slv-Shp: Sleeve-Shape; Slv-Len: Sleeve-Length. The $1^{st}/2^{nd}$ best results are highlighted in red/blue.

Methods \ Attributes	Category	Colour	Collar	Button	Pattern	Shape	Length	Slv-Shp	Slv-Len	**Average**
Imbalance ratio (1:x)[3]	2	138	210	242	476	2138	3401	4115	4162	
DDAN [7]	46.12	31.28	22.44	40.21	29.54	23.21	32.22	19.53	40.21	**31.64**
FashionNet [31]	48.45	36.82	25.27	43.85	31.60	27.37	38.56	20.53	45.16	**35.29**
DARN [22]	65.63	44.20	31.79	58.30	44.98	28.57	45.10	18.88	51.74	**43.24**
MTCT [10]	72.51	74.68	70.54	76.28	76.34	68.84	77.89	67.45	77.21	**73.53**
Over-Sampling* [24]	73.34	75.12	71.66	77.35	77.52	68.98	78.66	67.90	78.19	**74.30**
Down-Sampling* [34]	49.21	33.19	19.67	33.11	22.22	30.33	23.27	12.49	13.10	**26.29**
Cost-Sensitive* [20]	76.07	77.71	71.24	79.19	77.37	69.08	78.08	67.53	77.17	**74.49**
LMLE* [21]	75.90	77.62	70.84	78.67	77.83	71.27	79.14	69.83	80.83	**75.77**
CRL(C)*	**76.85**	**79.61**	**74.40**	**81.01**	**81.19**	**73.36**	**81.71**	**74.06**	**81.99**	**78.24**
CRL(I)*	**77.41**	**81.50**	**76.60**	**81.10**	**82.31**	**74.56**	**83.05**	**75.49**	**84.92**	**79.66**

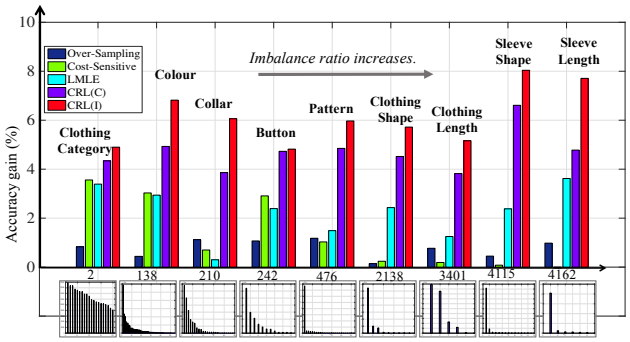

Figure 6. Model performance *additional* gain over the MTCT on 9 clothing attributes with increasing imbalance ratios on X-Domain.

baseline), along with the imbalance ratio for each clothing attribute. Figure 6 shows the comparisons and it is evident that CRL is clearly superior in learning severely imbalanced attributes, e.g. on "Sleeve Shape", CRL(C) and CRL(I) achieve 8% and 7% accuracy gain over MTCT respectively, as compared to the second best LMLE obtaining only 2% improvement. Qualitative examples are shown in Figure 7.

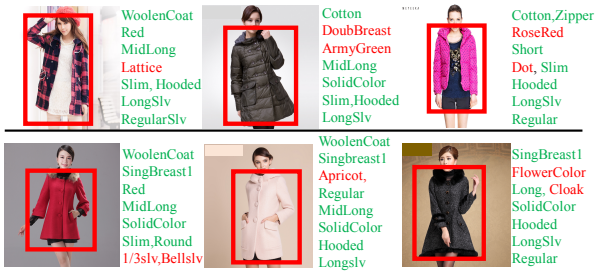

Figure 7. Examples of clothing attribute recognition by the CRL(I) model, with falsely predicted attributes in red.

3.3. Analysis on Rectification Loss and Hard Mining

We evaluated the effects of two different hard mining schemes (Class and Instance level) (Sec. 2.1), and three different CRL loss functions (Relative, Absolute, and Distribution comparisons) (Sec. 2.2). In total, we tested 6 different CRL variants. We evaluated the performance of these 6 CRL models by the accuracy gain over a non-imbalance learning baseline model: DeepID2 on CelebA and MTCT on X-domain. It is evident from Table 4 that: (1) All CRL models improve accuracy on both facial and clothing attribute recognition. (2) For both face and clothing, CRL(I+R) is the best and its performance advantage over other models is doubled on the more imbalanced X-Domain when compared to that on CelebA. (3) Most CRL models achieve greater performance gains on X-Domain than on CelebA. (4) Using the same loss function, instance-level hard mining is superior in most cases.

Table 4. Comparing different hard mining schemes (Class and Instance level) and loss functions (Relative(**R**), Absolute(**A**), and Distribution(**D**)). Metric: additional gain in average accuracy (%).

Dataset	CelebA			X-domain		
Loss function	A	R	D	A	R	D
Class Level	5.71	4.23	0.54	3.46	4.71	1.20
Instance Level	5.67	**5.85**	2.12	4.92	**6.13**	2.05

4. Conclusion

In this work, we formulated an end-to-end imbalanced deep learning framework for clothing and facial attribute recognition with very large scale imbalanced training data. The proposed Class Rectification Loss (CRL) model with batch-wise incremental hard positive and negative mining of the minority classes is designed to regularise deep model learning behaviour given training data with significantly imbalanced class distributions in very large scale data. Our experiments show clear advantages of the proposed CRL model over not only the state-of-the-art imbalanced data learning models but also dedicated attribute recognition methods for multi-label clothing and facial attribute recognition, surpassing the state-of-the-art LMLE model by 2% in average accuracy on the CelebA face benchmark and 4% on the more imbalanced X-Domain clothing benchmark, whilst having over threes time faster model training time advantage.

Acknowledgements

This work was partially supported by the China Scholarship Council, Vision Semantics Ltd., and the Royal Society Newton Advanced Fellowship Programme (NA150459).

References

[1] R. Akbani, S. Kwek, and N. Japkowicz. Applying support vector machines to imbalanced datasets. In *European Conference of Machine Learning*, pages 39–50, 2004. 1

[2] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005. 6, 7

[3] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 1

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. 1, 2

[5] C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data. *University of California, Berkeley*, pages 1–12, 2004. 2

[6] K. Chen, S. Gong, T. Xiang, and C. Loy. Cumulative attribute space for age and crowd density estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, Portland, Oregon, USA, 2013. 1

[7] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5315–5324, 2015. 1, 2, 5, 7, 8

[8] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 5

[9] Y. Cui, F. Zhou, Y. Lin, and S. Belongie. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[10] Q. Dong, S. Gong, and X. Zhu. Multi-task curriculum transfer deep learning of clothing attributes. In *IEEE Winter Conference on Applications of Computer Vision*, 2017. 2, 7, 8

[11] C. Drummond, R. C. Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. volume 11, 2003. 2, 6

[12] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 2

[13] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117, 2004. 6, 7

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 2

[15] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *ACM International Conference on Multimedia Retrieval*, page 153, 2014. 1

[16] F. Fernández-Navarro, C. Hervás-Martínez, and P. A. Gutiérrez. A dynamic over-sampling procedure based on sensitivity for multi-class problems. *Pattern Recognition*, 44(8):1821–1833, 2011. 6

[17] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person re-identification*, volume 1. Springer, 2014. 1

[18] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007. 2

[19] H. Han, W.-Y. Wang, and B.-H. 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

[20] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 1, 2, 6, 8

[21] C. Huang, Y. Li, C. Change Loy, and X. Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5375–5384, 2016. 1, 2, 3, 4, 5, 6, 8

[22] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *IEEE International Conference on Computer Vision*, 2015. 1, 2, 7, 8

[23] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv e-prints*, 2015. 5

[24] P. Jeatrakul, K. W. Wong, and C. C. 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. 1, 2, 6, 8

[25] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri. Cost sensitive learning of deep feature representations from imbalanced data. *arXiv e-prints*, 2015. 1, 2

[26] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009. 2

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 1

[28] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv e-prints*, 2013. 7

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 2

[30] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009. 4

[31] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016. 1, 2, 5, 7, 8

[32] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 1, 2, 5, 6, 7

[33] T. Maciejewski and J. Stefanowski. Local neighbourhood extension of smote for mining imbalanced data. In *IEEE International Conference on Data Mining*, pages 104–111, 2011. 1, 2

[34] I. Mani and I. Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, 2003. 6, 8

[35] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[36] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1717–1724, 2014. 1, 2

[37] 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. 2

[38] 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. 2, 6

[39] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 806–813, 2014. 1

[40] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3982–3991, 2015. 2

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1

[42] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, 2014. 6, 7

[43] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288, 2009. 1, 2

[44] K. M. Ting. A comparative study of cost-sensitive boosting algorithms. In *International Conference on Machine learning*, 2000. 1, 2

[45] E. Ustinova and V. Lempitsky. Learning deep embeddings with histogram loss. In *Advances in Neural Information Processing Systems*, 2016. 5

[46] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 2

[47] J. Wang, X. Zhu, S. Gong, and W. Li. Attribute recognition by joint recurrent learning of context and correlation. In *IEEE International Conference on Computer Vision*, 2017. 2

[48] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. In *IEEE International Conference on Computer Vision*, 2015. 2

[49] B. Zadrozny, J. Langford, and N. Abe. Cost-sensitive learning by cost-proportionate example weighting. In *IEEE International Conference on Data Mining*, pages 435–442, 2003. 2

[50] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014. 1, 2, 6

[51] Z.-H. Zhou and X.-Y. Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2006. 1, 2