M2m: Imbalanced Classification via Major-to-minor Translation

Jaehyung Kim* Jongheon Jeong* Jinwoo Shin
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea
{jaehyungkim, jongheonj, jinwoos}@kaist.ac.kr

Abstract

In most real-world scenarios, labeled training datasets are highly class-imbalanced, where deep neural networks suffer from generalizing to a balanced testing criterion. In this paper, we explore a novel yet simple way to alleviate this issue by augmenting less-frequent classes via translating samples (e.g., images) from more-frequent classes. This simple approach enables a classifier to learn more generalizable features of minority classes, by transferring and leveraging the diversity of the majority information. Our experimental results on a variety of class-imbalanced datasets show that the proposed method improves the generalization on minority classes significantly compared to other existing re-sampling or re-weighting methods. The performance of our method even surpasses those of previous state-of-the-art methods for the imbalanced classification.

1. Introduction

The recent success of deep neural networks (DNNs) across various computer vision problems [18, 38, 17, 37] has emerged due to the access to large-scale, annotated datasets collected from our visual world [40, 30, 1]. Despite having several well-organized datasets in research, *e.g.*, CIFAR [26] and ILSVRC [40], real-world datasets usually suffer from its expensive data acquisition process and the labeling cost. This commonly leads a dataset to have a "long-tailed" label distribution [34, 43]. Such *class-imbalanced* datasets make the standard training of DNN harder to generalize [44, 39, 9], particularly if one requires a class-balanced performance metric for a practical reason.

A natural approach in an attempt to bypass this *class-imbalance problem* is to re-balance the training objective artificially with respect to the class-wise sample sizes. Two of such methods are representative: (a) *re-weighting* the given loss function by a factor inversely proportional to the sample frequency in a class-wise manner [21, 25], and

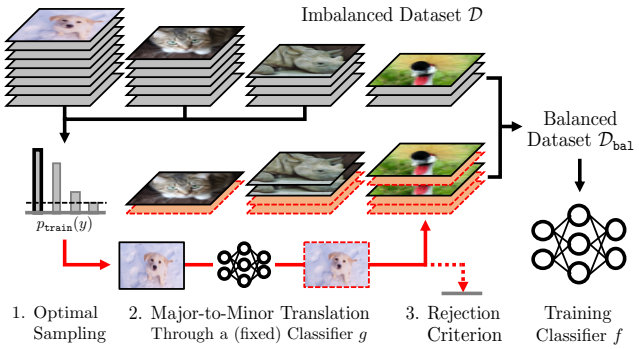


Figure 1. An overview of the proposed method, called *Major-to-minor Translation* (M2m). M2m is based on the over-sampling method, and attempts to replace the over-sampled (duplicated) minority samples with synthetic ones translated from other majority samples. The more details are presented in Section 2.

(b) *re-sampling* the given dataset so that the expected sampling distribution during training can be balanced, either by "over-sampling" the minority classes [23, 8] or "under-sampling" the majority classes [16].

However, naïvely re-balancing the objective usually results in harsh over-fitting to minority classes, since they cannot handle, the lack of minority information in essence. Several attempts have been made to alleviate this issue: Cui *et al*. [7] proposed the concept of "effective number" of samples as alternative weights in the re-weighting method. Cao *et al*. [4] found that both re-weighting and re-sampling can be much more effective when applied at the later stage of training, in case of neural networks. In the context of re-sampling, SMOTE [5] is a widely-used variant of the over-sampling method that mitigates the over-fitting via data augmentation, and several variants of SMOTE have been suggested accordingly [14, 15, 36]. A major drawback of these SMOTE-based methods is that they usually perform poorly when there exist only a few samples in the minority classes, *i.e.*, under regime of "extreme" imbalance, because they synthesize a new minority sample only using the existing samples of the same class.

Another line of research attempts to prevent the over-

*Equal contribution

fitting with a new regularization scheme that minority classes are more penalized, where the margin-based approaches generally suit well as a form of data-dependent regularizer [46, 9, 24, 4]. There have also been works that view the class-imbalance problem in the framework of active learning [10, 2] or meta-learning [44, 39, 41, 32].

Contribution. In this paper, we revisit the over-sampling framework and propose a new way of generating minority samples, coined *Major-to-minor Translation* (M2m). In contrast to other over-sampling methods, *e.g.*, SMOTE that applies data augmentation to minority samples to mitigate the over-fitting issue, we attempt to generate minority samples in a completely different way. The proposed M2m does *not* use the existing minority samples for the over-sampling. Instead, it use the *majority* (non-minority) samples and translate them to the target minority class using another classifier independently trained under the given imbalanced dataset. Our key finding is that, this method turns out to be very effective on learning more generalizable features in imbalanced learning: it does not overly use the minority samples, and leverages the richer information of the majority samples simultaneously.

Our minority over-sampling method consists of three components to improve the sampling quality. First, we propose an optimization objective for generating synthetic samples: a majority sample can be translated into a synthetic minority sample via optimizing it, while not affecting the performance of the majority class (even the sample is labeled to the minority class). Second, we design a sample rejection criterion based on the observation that generation from more majority class is more preferable. Third, based on the proposed rejection criterion, we suggest an optimal distribution for sampling a majority seed to be translated in our generation process.

We evaluate our method on various imbalanced classification problems, covering synthetically imbalanced datasets from CIFAR-10/100 [26] and ImageNet [32], and real-world imbalanced datasets including CelebA [31], SUN397 [45], Twitter [11] and Reuters [28] datasets. Despite its simplicity, our method significantly improves the balanced test accuracy compared to previous re-sampling or re-weighting methods across all the tested datasets. Our results even surpass those from LDAM [4], a current state-of-art margin-based method. Moreover, we found our method is particularly effective under "extreme" imbalance: in the case of Reuters of the most severe imbalance, we could improve the balanced accuracy by (relatively) 17.1% and 9.2% upon standard training and LDAM, respectively.

2. *M2m*: Major-to-minor translation

We consider a classification problem with K classes from a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, where $x \in \mathbb{R}^d$ and $y \in \{1, \cdots, K\}$ denote an input and the corresponding class label, respectively. Let $f : \mathbb{R}^d \to \mathbb{R}^K$ be a classifier designed to output K logits, which we want to train against the class-imbalanced dataset $\mathcal{D}$. We let $N := \sum_k N_k$ denote the total sample size of $\mathcal{D}$, where N_k is that of class k. Without loss of generality, we assume $N_1 \geq N_2 \geq \cdots \geq N_K$. In the *class-imbalanced* classification, the class-conditional data distributions $\mathcal{P}_k := p(x \mid y = k)$ are assumed to be invariant across training and test time, but they have different prior distributions, say $p_{\texttt{train}}(y)$ and $p_{\texttt{test}}(y)$, respectively: $p_{\texttt{train}}(y)$ is highly imbalanced while $p_{\texttt{test}}(y)$ is usually assumed to be the uniform distribution. The primary goal of the class-imbalanced learning is to train f from $\mathcal{D} \sim \mathcal{P}_{\texttt{train}}$ that generalizes well under $\mathcal{P}_{\texttt{test}}$ compared to the standard training, *e.g.*, empirical risk minimization (ERM) with an appropriate loss function $\mathcal{L}(f)$:

$$\min_f \ \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(f; x, y)]. \tag{1}$$

Our method is primarily based on over-sampling technique [23], a traditional and principled way to balance the class-imbalanced training objective via sampling minority classes more frequently. In other words, we assume a "virtually balanced" training dataset $\mathcal{D}_{\texttt{bal}}$ made from $\mathcal{D}$ such that the class k has $N_1 - N_k$ more samples, and the classifier f is trained on $\mathcal{D}_{\texttt{bal}}$ instead of $\mathcal{D}$.

A key challenge in over-sampling is to prevent *over-fitting* on minority classes, as the objective modified is essentially much biased to a few samples of minority classes. In contrast to most prior works that focus on performing data augmentation *directly* on minority samples to mitigate this issue [5, 32, 36], we attempt to augment minority samples in a completely different way: our method does *not* use the minority samples for the augmentation, but the majority samples.

2.1. Overview of M2m

Consider a scenario of training a neural network f on a class-imbalanced dataset $\mathcal{D}$. The proposed *Major-to-minor Translation* (M2m) attempts to construct a new balanced dataset $\mathcal{D}_{\texttt{bal}}$ for training f, by adding synthetic minority samples that are *translated* from other samples of (relatively) majority classes. There could be multiple ways to perform this "Major-to-minor" translation. In particular, a recent progress on cross-domain generation via generative adversarial networks [47, 6, 35] has made this more attractive, provided that much computational cost for additional training is acceptable. In this paper, on the other hand, we explore a much simpler and efficient approach: we translate a majority sample by optimizing it to *maximize* the target minority confidence of another baseline classifier g. Here, we assume the classifier g is a pre-trained neural network on $\mathcal{D}$ so that performs well (at least) on the training imbalanced dataset, *e.g.*, via standard ERM training. This implies

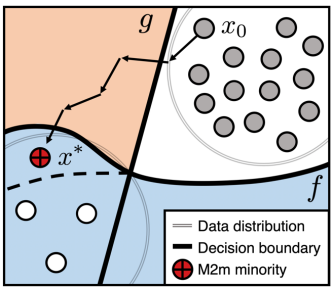


Figure 2. An illustration of M2m generation. A majority seed x_0 is translated to a synthetic minority x^* based on the decision boundary of g. By incorporating x^*, f learns an extended decision boundary of the target minority class.

that, g may be over-fitted to minority classes and does not necessarily generalize well under the balanced test dataset. We found this mild assumption on g is fairly enough to capture the information in the small minority classes and could generate surprisingly useful synthetic minority samples by utilizing the diversity of majority samples. On the other hand, f is the target network that we aim to train to perform well on the balanced testing criterion.

During the training of f, M2m utilizes the given classifier g to generate new minority samples, and the generated samples are added to $\mathcal{D}$ to construct $\mathcal{D}_{\texttt{bal}}$ on the fly. To obtain a single synthetic minority x^* of class k, our method solves an optimization problem starting from another training sample x_0 of a (relatively) major class $k_0 < k$:

$$x^* = \underset{x:=x_0+\delta}{\arg\min}\ \mathcal{L}(g; x, k) + \lambda \cdot f_{k_0}(x), \tag{2}$$

where $\mathcal{L}$ denotes the cross entropy loss and $\lambda > 0$ is a hyperparameter. In other words, our method "translates" a majority seed x_0 into x^*, so that g confidently classifies it as minority class k. The generated sample x^* is then labeled to class k and fed into f for training to perform better on $\mathcal{D}_{\texttt{bal}}$ and match the prediction of f to that of g. We do not force f in (2) to classify x^* to class k as well, but we restrict that f to have lower confidence on the original class k_0 by imposing a regularization term $\lambda \cdot f_{k_0}(x)$. Here, the regularization term $\lambda \cdot f_{k_0}(x)$ on the logit reduces the risk when x^* is labeled to k, whereas it may contain significant features of x_0 in the viewpoint of f. Intuitively, one can regard the overall process as teaching f to learn novel minority features which g considers it significant, *i.e.*, via extension of the decision boundary from the knowledge g. Figure 2 illustrates the basic idea of our method.

2.2. Underlying intuition on M2m

One may understand our method better by considering the case when g is an "oracle" (possibly the Bayes optimal) classifier, *e.g.*, (roughly) humans. Here, solving (2) essentially requires a transition of the original input x_0 of class k_0 with 100% confidence to another class k with respect to g: this would let g "erase and add" the features related to the class k_0 and k, respectively. Hence, in this case, our process corresponds to collecting more in-distribution minority data, which may be argued as the best way one could do to resolve the class-imbalance problem.

An intriguing point here is, however, that neural network models are very far from this ideal behavior, even when they achieve super-human performance. Instead, when f and g are neural networks, (2) often finds x^* that is very close to x_0, *i.e.*, similar to the phenomenon of *adversarial examples* [42, 12]. Nevertheless, we found our method still effectively improves the generalization of minority classes even in such cases. This observation is, in some sense, aligned to a recent claim that adversarial perturbation is not a "bug" in neural networks, but a "generalizable" feature [22].

In this paper, we hypothesize this counter-intuitive effectiveness of our method comes from mainly in two aspects: (a) the sample diversity in the majority dataset is utilized to prevent over-fitting on the minority classes, and (b) another classifier g is enough to capture the information in the small minority dataset. In this respect, adversarial examples from a majority to a minority can be regarded as one of natural ways to leverage the diverse features in majority examples useful to improve the generalization of the minority classes. It is also notable that our over-sampling method does not completely replace the existing dataset. Instead, our method only *augment* the minority classes, and our finding is that this augmentation turns out to be very effective than naïvely duplicating minority examples as done by the standard over-sampling. We further discuss a more detailed analysis to verify these claims, by performing an extensive ablation study in Section 3.4.

2.3. Detailed components of M2m

Sample rejection criterion. An important factor that affects the quality of the synthetic minority samples in our method is the quality of g, especially for g_{k_0}: a better g_{k_0} would more effectively "erase" important features of x_0 during the translation, thereby making the resulting minority samples more reliable. In practice, however, g is not that perfect: the synthetic samples still contain some discriminative features of the original class k_0, in which it may even harm the performance of f. This risk of "unreliable" generation becomes more harsh when N_{k_0} is small, as we assume that g is also trained on the given imbalanced dataset $\mathcal{D}$.

To alleviate this risk, we consider a simple criterion for *rejecting* each of the synthetic samples randomly with probability depending on k_0 and k:

$$\mathbb{P}(\text{Reject } x^* | k_0, k) := \beta^{(N_{k_0} - N_k)^+}, \tag{3}$$

where $(\cdot)^+ := \max(\cdot, 0)$, and $\beta \in [0, 1)$ is a hyperparameter which controls the reliability of g: the smaller β, the

Algorithm 1 Over-sampling via M2m

Input: A dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with $N = \sum_{k=1}^K N_k$. A classifier f. A pre-trained classifier g. $\lambda, \gamma, \eta, T > 0$ and $\beta \in [0, 1)$.

Output: A class-balanced dataset $\mathcal{D}_{\texttt{bal}}$

Initialize $\mathcal{D}_{\texttt{bal}} \leftarrow \mathcal{D}$
for $k = 2$ **to** K **do**
  $\Delta \leftarrow N_1 - N_k$
  for $i = 1$ **to** Δ **do**
    $k_0 \sim Q(k_0|k) \propto 1 - \beta^{(N_{k_0} - N_k)^+}$
    $x_0 \leftarrow$ A random sample of class k_0 in $\mathcal{D}$
    Initialize $x^* \leftarrow x_0 + \delta$ with a small noise δ
    for $t = 1$ **to** T **do**
      $\delta \leftarrow \nabla_{x^*}[\mathcal{L}(g; x^*, k) + \lambda \cdot f_{k_0}(x^*)]$
      $x^* \leftarrow x^* - \eta \cdot \frac{\delta}{||\delta||_2}$
    end for
    $R \sim \text{Bernoulli}(\beta^{(N_{k_0} - N_k)^+})$
    if $\mathcal{L}(g; x^*, k) > \gamma$ **or** $R = 1$ **then**
      $x^* \leftarrow$ A random sample of class k in $\mathcal{D}$
    end if
    $\mathcal{D}_{\texttt{bal}} \leftarrow \mathcal{D}_{\texttt{bal}} \cup \{(x^*, k)\}$
  end for
end for

more reliable g. For example, if $\beta = 0.999$, the synthetic samples are accepted with probability more than 99% if $N_{k_0} - N_k > 4602$. When $\beta = 0.9999$, on the other hand, it requires $N_{k_0} - N_k > 46049$ to achieve the same goal. This exponential modeling of the rejection probability is motivated by the *effective number* of samples [7], a heuristic recently proposed to model the observation that the impact of adding a single data point exponentially decreases at larger datasets. When a synthetic sample is rejected, we simply replace it by an existing minority sample from the original dataset $\mathcal{D}$ to obtain the balanced dataset $\mathcal{D}_{\texttt{bal}}$.

Optimal seed sampling. Another design choice of our method is *how to choose* a (majority) seed sample x_0 with class k_0 for each generation in (2). Based on the rejection criterion proposed in (3), we design a sampling distribution $Q(k_0|k)$ for selecting the class k_0 of initial point x_0 given target class k, by considering two aspects: (a) Q maximizes the *acceptance probability* $P_{\texttt{accept}}(k_0|k)$ under our rejection criterion, and (b) Q chooses *diverse* classes as much as possible, *i.e.*, the entropy $H(Q)$ is maximized. Namely, we are interested in the following optimization:

$$\max_Q \Big[\underbrace{\mathbb{E}_Q[\log P_{\texttt{accept}}]}_{\text{(a)}} + \underbrace{H(Q)}_{\text{(b)}} \Big]. \tag{4}$$

It is elementary to check that $Q = P_{\texttt{accept}}$ is the solution of the above optimization. Hence, due to the rejection probability (3), we choose:

$$Q(k_0|k) \propto 1 - \beta^{(N_{k_0} - N_k)^+}. \tag{5}$$

Once k_0 is selected, a sample x_0 is sampled uniformly at random among samples having the class k_0. The overall procedure of M2m is summarized in Algorithm 1.

Practical implementation via re-sampling. In practice of training a neural network f, *e.g.*, stochastic gradient descent (SGD) with a mini-batch sampling, M2m is implemented using a batch-wise re-sampling. More precisely, in order to simulate the generation of $N_1 - N_k$ samples for any $k = 1, 2, \cdots, K$, we perform the generation with probability $\frac{N_1 - N_{y_i}}{N_1} = 1 - N_{y_i}/N_1$, for all i in a given class-balanced mini-batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^m$.[1] For a single generation at index i, we first sample $k_0 \sim Q(k_0|y_i)$ following (5) until $k_0 \in \{y_i\}_{i=1}^m$, and select a seed x_0 of class k_0 randomly inside $\mathcal{B}$. Then, we solve the optimization (2) from x_0 toward class y_i via gradient descent for a fixed number of iterations T with a step size η. We accept the result sample x^* only if $\mathcal{L}(g; x^*, y_i)$ is less than $\gamma > 0$ for stability. Finally, if accepted, we replace (x_i, y_i) in $\mathcal{B}$ by (x^*, y_i).

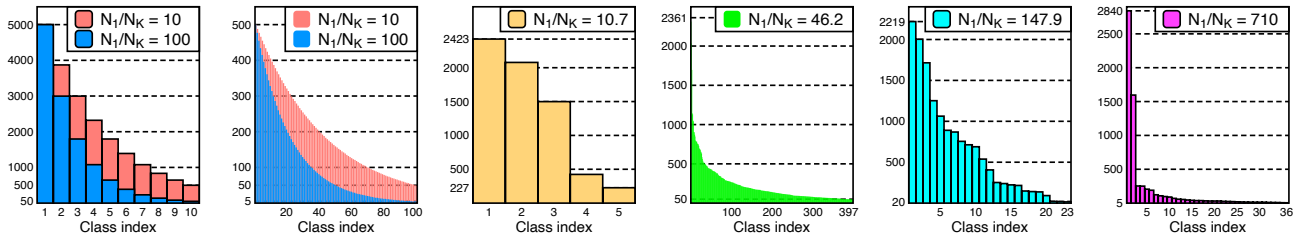
3. Experiments

We evaluate our method on various class-imbalanced classification tasks: synthetically-imbalanced variants of CIFAR-10/100 [26], ImageNet-LT[2] [32], CelebA [31], SUN397 [45], Twitter [11], and Reuters [28] datasets.[3] Figure 3 illustrates the class-wise sample distributions for the datasets considered in our experiments. The more details on the tested datasets are given in the supplementary material. To evaluate the classification performance of the models on the balanced test distribution, we mainly report two popular metrics: the *balanced accuracy* (bACC) [21, 44] and the *geometric mean scores* (GM) [27, 3], which are defined by the arithmetic and geometric mean over class-wise sensitivity (*i.e.*, recall), respectively. We remark that bACC is essentially equivalent to the standard accuracy metric for balanced datasets. All the values and error bars in this section are mean and standard deviation across three random trials, respectively. Overall, our results clearly demonstrate that minority synthesis via translating from majority consistently improves the efficiency of over-sampling, in terms of the significant improvement of the generalization in minority classes compared to other re-sampling baselines, across all the tested datasets. We also perform an ablation study to verify the effectiveness of our main ideas.

[1] Obtaining such a class-balanced mini-batch can be done via standard re-sampling.

[2] Results on ImageNet-LT can be found in the supplementary material.

[3] Code is available at https://github.com/alinlab/M2m



(a) CIFAR-LT-10 (b) CIFAR-LT-100 (c) CelebA-5 (d) SUN397 (e) Twitter (f) Reuters

Figure 3. An illustration of histograms on training sample sizes for the datasets used in this paper.

3.1. Experimental setup

Baseline methods. We consider a wide range of baseline methods, as listed in what follows: (*a*) empirical risk minimization (ERM): training on the cross-entropy loss without any re-balancing; (*b*) re-sampling (RS) [23]: balancing the objective from different sampling probability for each sample; (*c*) SMOTE [5]: a variant of re-sampling with data augmentation; (*d*) re-weighting (RW) [21]: balancing the objective from different weights on the sample-wise loss; (*e*) class-balanced re-weighting (CB-RW) [7]: a variant of re-weighting that uses the inverse of effective number for each class, defined as $(1-\beta^{N_k})/(1-\beta)$. Here, we use $\beta = 0.9999$; (*f*) deferred re-sampling (DRS) [4] and (*g*) deferred re-weighting (DRW) [4]: re-sampling and re-weighting is deferred until the later stage of the training, repsectively; (*h*) focal loss (Focal) [29]: the objective is up-weighted for relatively hard examples to focus more on the minority; (*i*) label-distribution-aware margin (LDAM) [4]: the classifier is trained to impose larger margin to minority classes. Roughly, the considered baselines can be classified into three categories: (i) "re-sampling" based methods - (*b, c, f*), (ii) "re-weighting" based methods - (*d, e, g*), and (iii) different loss functions - (*a, h, i*).

Training details. We train every model via stochastic gradient descent (SGD) with momentum of weight 0.9. The initial learning rate is set to 0.1, and "step decay" is performed during training where the exact scheduling across datasets is specified in the supplementary material. Although it did not affect much to our method, we also adopt the "linear warm-up" learning rate strategy [13] in the first 5 epochs, as the performance of some baseline methods, *e.g.*, re-weighting, highly depends on the use of this strategy. For CIFAR-10/100 and CelebA, we train ResNet-32 [19] for 200 epochs with mini-batch size 128, and set a weight decay of 2×10^{-4}. In case of SUN397, the pre-activation ResNet-18 model is used instead.[4] We ensure that all the input images are normalized over the training dataset, and have the size of 32×32 either by cropping or re-sizing, to be compatible with the given architectures. For Twitter and Reuters datasets, we train 2-layer fully-connected networks for 15 epochs with mini-batch size 64, and with a weight decay of 5×10^{-5}.

Details on M2m. When our method is applied, we use another classifier g of the same architecture to f that is pre-trained on the given (imbalanced) dataset via standard ERM training. Also, in a similar manner to that of [4], we use the deferred scheduling to our method, *i.e.*, we start to apply our method after the standard ERM training for a fixed number of epochs. The actual scheduling across datasets is specified in the supplementary material. We choose hyperparameters in our method from a fixed set of candidates, namely $\beta \in \{0.9, 0.99, 0.999\}$, $\lambda \in \{0.01, 0.1, 0.5\}$ and $\gamma \in \{0.9, 0.99\}$ based on the validation set. Unless otherwise stated, we fix $T = 10$ and $\eta = 0.1$ when performing a single generation step.

3.2. Long-tailed CIFAR datasets

We consider a "synthetically long-tailed" variant of CIFAR [26] datasets (CIFAR-LT-10/100) in order to evaluate our method on various levels of imbalance, where the original datasets are class-balanced. To simulate the long-tailed distribution frequently appeared in imbalanced datasets, we control the *imbalance ratio* $\rho > 1$ and artificially reduce the training sample sizes of each class except the first class, so that: (a) N_1/N_K equals to ρ, and (b) N_k in between N_1 and N_K follows an exponential decay across k. We keep the test dataset unchanged during this process, *i.e.*, it is still perfectly balanced, thereby measuring accuracy on this dataset is equivalent to measuring the balanced accuracy. We consider two imbalance ratios $\rho \in \{100, 10\}$ each for CIFAR-LT-10 and 100. See Figure 3(a) and 3(b) for a detailed illustration of the sample distribution.

Table 1 summarizes the main results. In overall, the results show that our method consistently improves the bACC by a large margin, across all the tested baselines. These results even surpass the "LDAM+DRW" baseline [4], which is known to be the state-of-the-art to the best of our knowledge. Moreover, we point out, in most cases, our method could further improve bACC when applied upon the LDAM

[4] We remark this model is larger than ResNet-32 used for CIFAR and CelebA datasets, as it has roughly 4× more channels.

Dataset		CIFAR-LT-10				CIFAR-LT-100			
Imbalance ratio		$N_1/N_K = 100$		$N_1/N_K = 10$		$N_1/N_K = 100$		$N_1/N_K = 10$	
Loss	Re-balancing	bACC	GM	bACC	GM	bACC	GM	bACC	GM
ERM	-	68.7$_{\pm 1.43}$	66.4$_{\pm 1.69}$	86.0$_{\pm 0.69}$	85.8$_{\pm 0.50}$	37.2$_{\pm 1.12}$	21.5$_{\pm 1.66}$	56.2$_{\pm 0.69}$	51.8$_{\pm 0.63}$
ERM	RS	70.4$_{\pm 1.15}$	69.0$_{\pm 1.36}$	86.6$_{\pm 0.37}$	86.4$_{\pm 0.37}$	31.6$_{\pm 1.26}$	17.7$_{\pm 1.33}$	54.8$_{\pm 0.47}$	50.3$_{\pm 0.68}$
ERM	SMOTE	71.5$_{\pm 0.57}$	70.2$_{\pm 0.93}$	85.7$_{\pm 0.25}$	85.5$_{\pm 0.26}$	34.0$_{\pm 0.33}$	19.6$_{\pm 0.36}$	53.8$_{\pm 0.93}$	49.4$_{\pm 1.15}$
ERM	RW	72.8$_{\pm 0.33}$	72.0$_{\pm 0.29}$	86.6$_{\pm 0.18}$	86.5$_{\pm 0.16}$	30.1$_{\pm 0.59}$	17.6$_{\pm 0.85}$	56.0$_{\pm 0.35}$	52.0$_{\pm 0.51}$
ERM	CB-RW	71.2$_{\pm 1.14}$	70.0$_{\pm 1.28}$	86.8$_{\pm 0.49}$	86.6$_{\pm 0.53}$	38.6$_{\pm 0.46}$	22.5$_{\pm 0.49}$	55.9$_{\pm 0.24}$	52.0$_{\pm 0.42}$
ERM	DRS	75.2$_{\pm 0.26}$	73.9$_{\pm 0.17}$	87.1$_{\pm 0.26}$	87.0$_{\pm 0.29}$	41.5$_{\pm 0.21}$	31.0$_{\pm 0.21}$	57.7$_{\pm 0.40}$	54.8$_{\pm 0.33}$
ERM	**M2m (ours)**	**78.3**$_{\pm 0.16}$	**77.8**$_{\pm 0.16}$	**87.9**$_{\pm 0.21}$	**87.5**$_{\pm 0.15}$	**42.9**$_{\pm 0.16}$	**33.0**$_{\pm 0.11}$	**58.2**$_{\pm 0.08}$	**55.3**$_{\pm 0.05}$
Focal	-	68.3$_{\pm 1.19}$	65.5$_{\pm 1.71}$	85.3$_{\pm 0.47}$	85.1$_{\pm 0.47}$	37.7$_{\pm 1.38}$	22.1$_{\pm 1.49}$	55.3$_{\pm 0.42}$	50.7$_{\pm 0.43}$
LDAM	-	72.8$_{\pm 0.37}$	70.8$_{\pm 0.65}$	86.2$_{\pm 0.12}$	86.0$_{\pm 0.15}$	39.5$_{\pm 0.69}$	20.8$_{\pm 0.49}$	54.7$_{\pm 0.16}$	44.1$_{\pm 0.53}$
LDAM	DRW	77.1$_{\pm 0.49}$	76.7$_{\pm 0.59}$	87.1$_{\pm 0.28}$	86.9$_{\pm 0.28}$	42.1$_{\pm 0.09}$	29.2$_{\pm 0.27}$	56.9$_{\pm 0.15}$	50.4$_{\pm 0.29}$
LDAM	**M2m (ours)**	**79.1**$_{\pm 0.19}$	**78.6**$_{\pm 0.19}$	**87.5**$_{\pm 0.15}$	**87.4**$_{\pm 0.19}$	**43.5**$_{\pm 0.22}$	**34.2**$_{\pm 0.62}$	**57.6**$_{\pm 0.14}$	**51.8**$_{\pm 0.38}$

Table 1. Comparison of classification performance on the four different types of long-tailed CIFAR-10/100 datasets.

Datasets		CelebA-5		SUN397		Twitter		Reuters	
Imbalance ratio		$N_1/N_K \approx 10.7$		$N_1/N_K \approx 46.2$		$N_1/N_K \approx 147.9$		$N_1/N_K = 710$	
Loss	Re-balancing	bACC	GM	bACC	GM	bACC	GM	bACC	GM
ERM	-	72.7$_{\pm 1.24}$	69.4$_{\pm 0.97}$	31.5$_{\pm 0.07}$	20.2$_{\pm 0.74}$	74.7$_{\pm 0.46}$	65.2$_{\pm 1.10}$	59.8$_{\pm 1.17}$	53.8$_{\pm 1.75}$
ERM	RS	72.5$_{\pm 0.93}$	70.4$_{\pm 1.37}$	28.4$_{\pm 0.19}$	19.8$_{\pm 1.10}$	75.8$_{\pm 0.30}$	70.4$_{\pm 1.67}$	63.3$_{\pm 0.90}$	57.4$_{\pm 1.03}$
ERM	SMOTE	72.8$_{\pm 1.07}$	70.7$_{\pm 0.84}$	23.7$_{\pm 0.09}$	14.8$_{\pm 0.39}$	75.8$_{\pm 0.38}$	69.5$_{\pm 0.30}$	62.5$_{\pm 1.30}$	56.8$_{\pm 1.69}$
ERM	RW	74.5$_{\pm 0.50}$	73.4$_{\pm 0.87}$	31.3$_{\pm 0.20}$	25.3$_{\pm 0.12}$	76.2$_{\pm 0.95}$	73.5$_{\pm 1.46}$	65.0$_{\pm 1.08}$	59.2$_{\pm 1.84}$
ERM	CB-RW	74.2$_{\pm 0.59}$	72.3$_{\pm 0.50}$	31.7$_{\pm 0.13}$	25.1$_{\pm 0.51}$	77.5$_{\pm 0.40}$	73.6$_{\pm 0.79}$	64.8$_{\pm 0.45}$	57.6$_{\pm 1.62}$
ERM	DRS	73.1$_{\pm 0.68}$	71.2$_{\pm 0.62}$	30.7$_{\pm 0.34}$	24.2$_{\pm 0.40}$	77.8$_{\pm 0.85}$	74.3$_{\pm 1.48}$	62.4$_{\pm 0.39}$	56.0$_{\pm 1.34}$
ERM	**M2m (ours)**	**75.6**$_{\pm 0.16}$	**74.6**$_{\pm 0.34}$	**32.4**$_{\pm 0.17}$	**25.8**$_{\pm 0.29}$	**78.2**$_{\pm 0.35}$	**74.8**$_{\pm 0.78}$	**66.3**$_{\pm 0.42}$	**60.5**$_{\pm 0.52}$
Focal	-	72.7$_{\pm 0.57}$	69.7$_{\pm 1.42}$	31.2$_{\pm 0.14}$	21.3$_{\pm 0.71}$	74.2$_{\pm 2.35}$	70.4$_{\pm 4.03}$	59.4$_{\pm 0.42}$	53.0$_{\pm 0.74}$
LDAM	-	73.0$_{\pm 1.14}$	68.0$_{\pm 2.19}$	30.2$_{\pm 0.10}$	14.4$_{\pm 0.83}$	74.6$_{\pm 0.40}$	66.1$_{\pm 2.28}$	63.0$_{\pm 1.36}$	57.6$_{\pm 0.50}$
LDAM	DRW	74.4$_{\pm 0.33}$	72.3$_{\pm 0.82}$	31.6$_{\pm 0.10}$	23.6$_{\pm 0.36}$	78.0$_{\pm 0.87}$	74.4$_{\pm 1.28}$	64.1$_{\pm 0.31}$	56.9$_{\pm 1.08}$
LDAM	**M2m (ours)**	**75.9**$_{\pm 1.09}$	**75.0**$_{\pm 0.94}$	**33.3**$_{\pm 0.20}$	**24.9**$_{\pm 0.76}$	**78.8**$_{\pm 0.21}$	**76.0**$_{\pm 0.23}$	**70.0**$_{\pm 0.68}$	**63.9**$_{\pm 0.49}$

Table 2. Comparison of classification performance on the four naturally imbalanced datasets: CelebA-5, SUN397, Twitter and Reuters. In case of Reuters, η is adjusted to 1.0 when training M2m models regarding the numerical range of the dataset.

training scheme (see "LDAM+AMO"): this indicates that the performance gain from our method is fairly orthogonal to that of LDAM, *i.e.*, the margin-based approach, which suggests a new promising direction of improving the generalization when a neural network model suffers from a problem of small data.

3.3. Real-world imbalanced datasets

We further verify the effectiveness of M2m on four well-known, *naturally* imbalanced datasets: CelebA [31], SUN397 [45], Twitter [11] and Reuters [28] datasets. More detailed information for each of these datasets is demonstrated in Figure 3 and the supplementary material.

CelebA is originally a multi-labeled dataset, and we port this to a 5-way classification task by filtering only the samples with five non-overlapping labels about hair colors. We also subsampled the full dataset by $1/20$ while maintaining the imbalance ratio $\rho \approx 10.7$, in attempt to make the task more difficult. We denote the resulting dataset by CelebA-5.

Although Twitter and Reuters datasets are from natural language processing, we also evaluate our method on them to test the effectiveness under much extreme imbalance. Here, we remark that the imbalance ratio N_1/N_k of these two datasets are about 150 and 710, respectively, which are much higher than the other image datasets tested. In case of Reuters, we exclude the classes having less than 5 samples in the test set for more reliable evaluation, resulting a dataset of 36 classes.

Table 2 shows the results. Again, M2m performs best amongst other baseline methods, demonstrating the effectiveness of our method under natural imbalance, as well as wider applicability of our algorithm beyond image clas-

# Seeds	bACC (Δ)	GM (Δ)
10	74.9$_{\pm 0.29}$ (-4.34%)	73.7$_{\pm 0.33}$ (-5.27%)
50	76.2$_{\pm 0.30}$ (-2.68%)	75.3$_{\pm 0.29}$ (-3.21%)
100	76.5$_{\pm 0.34}$ (-2.30%)	75.6$_{\pm 0.41}$ (-2.83%)
200	76.7$_{\pm 0.51}$ (-2.04%)	75.9$_{\pm 0.59}$ (-2.44%)
500	77.4$_{\pm 0.38}$ (-1.15%)	76.8$_{\pm 0.31}$ (-1.29%)
Full	**78.3**$_{\pm \mathbf{0.16}}$ (-0.00%)	**77.8**$_{\pm 0.16}$ (-0.00%)

Table 3. Comparison of classification performance across various number of samples allowed to be a seed sample x_0. Δ indicates the relative gap from the original result presented in "Full".

Methods	Major (2)	Minor (8)	bACC	GM
M2m ($\lambda = 0$)	92.8$_{\pm 0.97}$	73.0$_{\pm 0.10}$	76.9$_{\pm 0.15}$	76.5$_{\pm 0.11}$
M2m-Clean	78.4$_{\pm 2.45}$	72.7$_{\pm 0.60}$	73.5$_{\pm 0.81}$	73.0$_{\pm 0.93}$
ERM-RS	92.8$_{\pm 1.50}$	64.8$_{\pm 1.18}$	70.4$_{\pm 1.15}$	69.0$_{\pm 1.36}$
M2m-RS	92.9$_{\pm 2.99}$	69.4$_{\pm 0.84}$	74.1$_{\pm 0.10}$	73.1$_{\pm 0.14}$
M2m-RS-Rand	93.6$_{\pm 2.34}$	66.1$_{\pm 1.04}$	71.6$_{\pm 0.36}$	70.3$_{\pm 0.80}$
M2m	**93.3**$_{\pm 0.85}$	**74.6**$_{\pm 0.34}$	**78.3**$_{\pm 0.16}$	**77.8**$_{\pm 0.16}$

Table 4. Comparison of classification performance across various types of ablations. We report the number of majority and minority classes in the parentheses.

sification. Remarkably, the significant results on Reuters dataset compared to the others suggest that our method can be even more effective under a regime of "extremely" imbalanced datasets, as Reuters has a much larger imbalance ratio than the others.

3.4. Ablation study

We conduct an extensive ablation study to present a detailed analysis of the proposed method. All the experiments in this section are performed with ResNet-32 models, trained on CIFAR-LT-10 with the imbalance ratio $\rho = 100$. We additionally report the balanced test accuracy over *majority* and *minority* classes, namely "Major" and "Minor" respectively, to further identify the relative impacts on those two classes separately. We divide the whole classes into "majority" and "minority" classes, so that the majority classes consist of top-k frequent classes with respect to the training set where k is the minimum number that $\sum_k N_k$ exceeds 50% of the total. We denote the minority classes as the remaining classes. We provide more discussion in the supplementary material.

Diversity on seed samples. In Section 2.1, we hypothesize that the effectiveness of our method mainly comes from utilizing a much diversity in the majority samples to prevent the over-fitting to the minority classes. To verify this, we consider an ablation that the candidates of "seed samples" are limited: more concretely, we control the size of seed sample pools per each class to a fixed subset of the training set, made before training f. In Table 3, the accuracy of minority classes is progressively increased as seed sample pools become diverse. This clear trend indicates that M2m makes use of the diversity of majority classes for preventing the over-fitting to the minority classes.

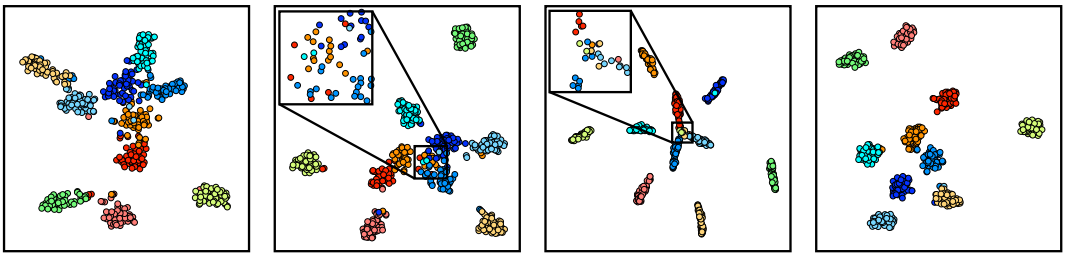
The effect of λ. In the optimization objective (2) for the generation step in M2m, we impose a regularization term $\lambda \cdot f_{k_0}(x)$ to improve the quality of synthetic samples: they might confuse f if themselves still contain important features of the original class in a viewpoint of f. To verify the effect of this term, we consider an ablation that λ is set to 0, and compare the performance to the original method. As reported in Table 4, we found a certain level of degradation in the balanced test accuracy at this ablation, which shows the effectiveness of the proposed regularization.

Over-sampling from the scratch. As specified in Section 3.1, we use the "deferred" scheduling to our method by default, *i.e.*, we start to apply our method after the standard ERM training for a fixed number of epochs. We have also considered a simple ablation where this strategy is not used, namely "M2m-RS". The results in Table 4 show that M2m-RS still outperforms any other baselines (reported in Table 1) except the ones that the deferred scheduling is used, *i.e.*, DRS and DRW, and this further verifies the effectiveness of our method.

Labeling as a targeted class. Our primary assumption on the pre-trained classifier g does not require that g itself to generalize well on the minority classes (see Section 2.1). This implies that solving (2) with g may not end up with a synthetic sample that contains generalizable features of the target minority class. To examine how much the generated samples would be correlated to the target classes, we consider another ablation upon M2m-RS:[5] instead of labeling the generated sample as the target class, the ablated method "M2m-RS-Rand" labels it to a "random" class chosen from all the possible classes (except for the target and original classes). The results shown in Table 4 indicate that M2m-RS-Rand generalizes much worse than its counterpart M2m-RS on the minority classes, which indeed confirms that the correctly-labeled synthetic samples could improve the generalization of the minority classes.

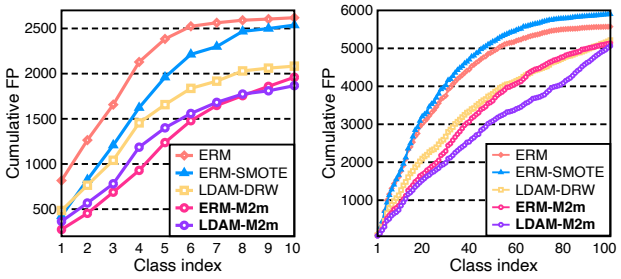
Comparison of t-SNE embeddings. To further validate the effectiveness of our method, we visualize and compare the penultimate features learned from various training methods (including ours) using t-SNE [33]. Each embedding is computed from a randomly-chosen subset of training samples in the CIFAR-LT-10 ($\rho = 100$), so that it consists of 50 samples per each class. Figure 4 illustrates the results, and shows that the embedding from our training method (M2m) is of much separable features compared to other methods: one could successfully distinguish each cluster

[5] Here, we attempt to opt out any potential effect from using DRS, for more clearer evaluation.



(a) ERM (b) ERM-SMOTE (c) LDAM-DRW (d) ERM-M2m (ours)

Figure 4. Visualization of the penultimate features via t-SNE computed from a balanced subset of CIFAR-LT-10 with ResNet-32.



(a) CIFAR-LT-10 (b) CIFAR-LT-100

Figure 5. Comparisons on cumulative number of false positive samples across class indices ($\sum_k \mathrm{FP}_k$) on CIFAR-LT-10/100 test set in ResNet-32. Each plot reaches the total number of mistakes, *i.e.*, the sum of off-diagonal entries in the confusion matrix.

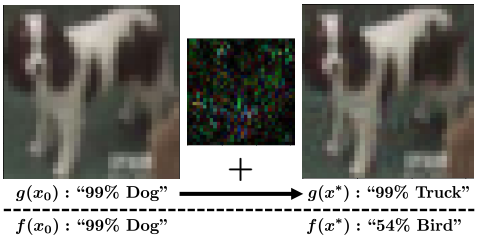


Figure 6. An illustration of a synthetic minority sample by M2m, where g is assumed to be ResNet-32 trained by standard ERM. The noise image is amplified by 10 for better visibility.

under the M2m embedding (even though they are from minority classes), while others have some obscure region.

Comparison of cumulative false positive. In Figure 5, we plot how the number of false positive (FP) samples increases as summed over classes, namely $\sum_k \mathrm{FP}_k$, from the most frequent class to the least one. Here, FP_k indicates the number of misclassified samples by predicting them to class k in the test set. We compute each plot with the *balanced* test sets of CIFAR-LT-10/100, thereby a well-trained classifier would show a plot close to linear: it indicates the classifier mistakes more evenly over the classes. Overall, one could see that the curve made by our method consistently below the others with much linearity. This implies our method makes less false positives, and even better, they are more uniformly distributed over the classes. This is a desirable property in the context of imbalanced learning.

The use of adversarial examples. As mentioned in Section 2.2, the generation under M2m often ends up with a synthetic minority sample that is very close to the original (before translation) as like the *adversarial example*. This indeed happens when f and g are neural networks as assumed here, *i.e.*, ResNet-32, as illustrated in Figure 6. To understand more on how such adversarial perturbations affect our method, we consider a simple ablation, which we call "M2m-Clean": recall that our method synthesizes a minority sample x^* from a seed majority sample x_0. This ablation uses the "clean" x_0 instead of x^* for over-sampling. Under the identical training setup, we notice a significant reduction in the balanced accuracy of M2m-Clean compared to the original M2m (see Table 4). This observation reveals that the adversarial perturbations ablated are extremely crucial to make our method to work, regardless of a small noise.

4. Conclusion

We propose a new over-sampling method for imbalanced classification, called *Major-to-minor Translation* (M2m). We found the diversity in majority samples could much help the class-imbalanced training, even with a simple translation method using a pre-trained classifier. This suggests a promising way to overcome the long-standing class-imbalance problem, and exploring more powerful methods to perform this Major-to-minor translation, *e.g.*, CycleGAN [47], would be an interesting future research. The problems we explored in this paper also lead us to an essential question that whether an adversarial perturbation could be a good feature. Our findings suggest that it could be, at least for the purpose of imbalanced learning, where the minority classes suffer over-fitting due to insufficient data. We believe our method could open up a new direction of research both in imbalanced learning and adversarial examples.

Acknowledgements

This work was supported by Samsung Electronics and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)).

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8M: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1

[2] Josh Attenberg and Seyda Ertekin. Class imbalance and active learning. *Imbalanced Learning: Foundations, Algorithms, and Applications*, 2013. 2

[3] Paula Branco, Luís Torgo, and Rita P Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys (CSUR)*, 49(2):31, 2016. 4

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1, 2, 5, 11, 12

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002. 1, 2, 5

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, 2018. 2

[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 4, 5, 11

[8] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[9] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 1, 2

[10] Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. Learning on the border: active learning in imbalanced data classification. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, 2007. 2

[11] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2011. 2, 4, 6, 11

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 3

[13] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 5

[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 (ICIC)*, pages 878–887. Springer, 2005. 1

[15] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1322–1328. IEEE, 2008. 1

[16] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2008. 1

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 1

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 12

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 5, 11

[20] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2016. 11

[21] 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 (CVPR)*, 2016. 1, 4, 5

[22] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[23] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proceedings of the International Conference on Artificial Intelligence (ICAI)*, 2000. 1, 2, 5

[24] 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 (CVPR)*, 2019. 2

[25] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2017. 1

[26] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Department of Computer Science, University of Toronto, 2009. 1, 2, 4, 5, 11

[27] M Kubat. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 179–186, 1997. 4

[28] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 2004. 2, 4, 6, 11

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 5

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014. 1

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 2, 4, 6, 11

[32] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4, 12

[33] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008. 7

[34] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[35] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. InstaGAN: Instance-aware image-to-image translation. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[36] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 11

[37] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017. 1

[38] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[39] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018. 1, 2

[40] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 1, 12

[41] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[42] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 3

[43] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[44] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1, 2, 4

[45] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 2, 4, 6, 11

[46] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 2, 8

Supplementary Material

M2m: Imbalanced Classification via Major-to-minor Translation

A. Details on the datasets

CIFAR-LT-10/100. CIFAR-10/100 datasets [26] consist of 60,000 RGB images of size 32×32, 50,000 for training and 10,000 for testing. Each image in the two datasets is corresponded to one of 10 and 100 classes, respectively. In our experiments, we construct "synthetically long-tailed" variants of CIFAR-10/100, namely CIFAR-LT-10/100, respectively [4]. We hold-out 10% of the test set to construct a validation set, and use the remaining for testing. We use ResNet-32 [19] with a mini-batch size 128, and set a weight decay of 2×10^{-4}. We train the network for 200 epochs with an initial learning rate of 0.1. We follow the learning rate schedule used by [7] for fair comparison: the initial learning rate is set to 0.1, and we decay it by a factor of 100 at 160-th and 180-th epoch. When the deferred scheduling [4] is used, *e.g.*, DRS, DRW and our method, it is applied after 160 epochs of standard training.

CelebA-5. CelebFaces Attributes (CelebA) dataset [31] is a multi-labeled face attributes dataset. It is originally composed of 202,599 number of RGB face images with 40 binary attributes annotations per image. We port this CelebA to a 5-way classification task by filtering only the samples with five non-overlapping labels about hair colors: namely, "blonde", "black", "bald", "brown", and "gray". This is in a similar manner as done in [36]. We denote the resulting dataset by CelebA-5. We pick out 50 and 100 samples per each class for validation and testing. We use ResNet-32 [19] with a mini-batch size 128, and set a weight decay of 2×10^{-4}. We train the network for 90 epochs with an initial learning rate of 0.1. We decay the learning rate by 0.1 at epoch 30 and 60. When the deferred scheduling is used, it is applied after 60 epochs of standard training.

SUN397. Scene UNderstanding (SUN) [45] is a dataset for a scene categorization. It originally consists of 108,754 RGB images which are labeled with 397 classes. For the inputs, center patches are first extracted and they are resized to 32×32. We hold-out 10 and 40 samples per each class for validation and testing, respectively, as the dataset itself does not provide any separated split for testing. We use pre-activation ResNet-18 [19] which roughly has $4\times$ more channels with a mini-batch size 128, and set a weight decay of 2×10^{-4}. We train the network for 90 epochs with an initial learning rate of 0.1. We decay the learning rate by 0.1 at epoch 30 and 60. When the deferred scheduling is used, it is applied after 60 epochs of standard training.

Twitter. Twitter [11] is a dataset for a part-of-speech (POS) tagging task in social media text with 25 classes. Each sample is a pair of a token and a tag, *e.g.*, "(books, common noun)" and "(#acl, hashtag)", where each token is embedded into a 50-dimensional vector via a pre-defined word-embedding [20]. We discarded two classes with zero test samples and obtained 14,614 training samples with 23 classes. We use 2-layer fully-connected network with a hidden layer size of 256 and a ReLU nonlinearity. We set a mini-batch size 64 and a weight decay of 5×10^{-5}. We train the network for 15 epochs with an initial learning rate 0.1 and decay the learning rate by 0.1 at epoch 10. When the deferred scheduling is used, it is applied after 10 epochs of standard training.

Reuters. Reuters [28] is a dataset for a text categorization task which predicts the subject of a given text. As an input, 1000-dimensional bag-of-words vectors are given, which are processed from a news story document. It is originally composed of 52 classes, but we discarded the classes that have less than 5 test samples for a reliable evaluation, obtaining a subset of the full dataset of 36 classes with 6436 training samples. We hold-out 10% of training samples to construct a validation set. We use 2-layer fully-connected network with a hidden layer size of 256 and a ReLU nonlinearity. We set a mini-batch size 64 and a weight decay of 5×10^{-5}. We train the network for 15 epochs with an initial learning rate 0.1 and decay the learning rate by 0.1 at epoch 10. When the deferred scheduling is used, it is applied after 10 epochs of standard training.

B. More results from ablation study

Generation from another classifier g. As mentioned, our method introduces another classifier g to generate synthetic minority x^* independently from the training classifier f. This is because using f itself instead of g in the optimization objective (2) would let the synthetic samples already confident in the target minority class to f, and this makes the overall training process redundant. To further validate the importance of using g, we consider an ablation called "M2m-Self": instead of using g, "M2m-Self" uses f for generating minority samples. As reported in Table 5, one could immediately see that M2m-Self only shows marginal improvement from DRS, which is much inferior than the original M2m.

Methods	bACC (Δ)	GM (Δ)
ERM-DRS	$75.2_{\pm 0.26}$ (-3.96%)	$73.9_{\pm 0.32}$ (-5.01%)
M2m-Self	$75.9_{\pm 0.27}$ (-3.07%)	$74.9_{\pm 0.32}$ (-3.73%)
M2m-No-Reject	$77.4_{\pm 0.33}$ (-1.15%)	$76.8_{\pm 0.40}$ (-1.29%)
M2m ($\gamma = 0$)	$76.9_{\pm 0.19}$ (-1.79%)	$76.4_{\pm 0.20}$ (-1.80%)
M2m	$78.3_{\pm 0.16}$ (-0.00%)	$77.8_{\pm 0.16}$ (-0.00%)
M2m-Ensemble	$78.5_{\pm 0.20}$ (+0.26%)	$78.0_{\pm 0.22}$ (+0.26%)

Table 5. Comparison of classification performance across various types of ablations. Δ indicates the relative gap from the original result presented in "M2m". All the values and error bars are mean and standard deviation across three random trials, respectively.

Loss	Re-balancing	bACC (Δ)	GM (Δ)
ERM	-	$38.6_{\pm 0.75}$	$26.9_{\pm 0.78}$
ERM	DRS	$40.8_{\pm 0.67}$	$31.6_{\pm 1.05}$
ERM	**M2m (ours)**	$\mathbf{42.2}_{\pm 0.51}$	$\mathbf{33.1}_{\pm 0.64}$
LDAM	-	$41.0_{\pm 0.07}$	$28.5_{\pm 0.11}$
LDAM	DRW	$43.0_{\pm 0.17}$	$34.5_{\pm 0.17}$
LDAM	**M2m (ours)**	$\mathbf{43.7}_{\pm 0.26}$	$\mathbf{35.1}_{\pm 0.35}$

Table 6. Comparison of classification performance on ImageNet-LT. All the values and error bars are mean and standard deviation across three random trials, respectively.

Using multiple classifiers for generation. Since our method is not restricted to use the only one pre-trained classifier g in the optimization (2), the multiple classifiers g_i for $i = 1, \dots, m$ can be used to improve the quality of generation. To verify the additional gain from multiple classifiers, we consider an ablation called "M2m-Ensemble": use the ensemble of the classifiers ($m = 2$) for generation instead of the single classifier. Here, we use the same architecture ResNet-32 for g_1 and g_2 and use a higher γ due to the smoothed prediction from the ensemble. The results in Table 5 show that M2m-Ensemble slightly perform better than M2m. It indicates that our method can benefit from the stronger classifier.

Rejection criteria. We also propose a sample rejection criteria to alleviate the risk of unreliable generation, possibly due to a weak generalization of g. To verify the effect of this rejection criteria, we consider an ablation, namely "M2m-No-Reject", which does not use this rejection policy in training. In other words, all the generated samples are used to train f. The results in Table 5 show that M2m-No-Reject performs significantly worse than M2m. This indeed confirms the gain from using the proposed rejection criteria.

The effect of γ. As specified in Algorithm 1 in the main paper, we set a threshold γ to filter out the synthetic samples which the generation objective is not sufficiently minimized, mainly due to the limited budget. To evaluate the practical effectiveness of using γ, here we consider an ablation that this thresholding is not used, equivalently when $\gamma = \infty$. As reported in Table 5, we indeed observe a performance degradation by not using γ. This reveals that the confidence level in g affects the final quality of the generation.

C. Results on ImageNet-LT

We additionally evaluate our method on ImageNet-LT [32] dataset, a subset of ImageNet dataset [40] with a synthetic imbalance following the Pareto distribution of the power $\alpha = 6$. It is composed of 115,846 training samples with 1,000 categories, 1,280 images in the maximal class and 5 images in the minimal class. A more detailed distribution is presented in Figure 7. We use the randomly-resized cropping and the horizontal flipping as a data augmentation, and all the images are resized to 128×128. We hold-out 20 samples per class randomly from the original ImageNet training set to form a validation set, and the original (roughly balanced) ImageNet validation set is used for testing. We use ResNet-50 [18] with a mini-batch size 256 and set a weight decay of 10^{-4}. We train the network for 200 epochs with an initial learning rate of 0.1 and it is decayed by 0.1 at epoch 160 and 180. When the deferred scheduling is used, *e.g.*, DRS, DRW and our method, it is applied after 160 epochs of standard training. We evaluate our method with followings which show the best performance among the baselines in the experiments in the main paper: (*a*) ERM-DRS and (*b*) LDAM-DRW [4]. We report the *balanced accuracy* (bACC) and the *geometric mean scores* (GM). As reported in Table 6, our method, M2m, significantly outperforms the baselines. In the case of ERM loss, compare to DRS, M2m shows 3.43 % and 4.75 % relative gains in bACC and GM, respectively. Furthermore, with a margin-based loss function LDAM, the improvement is much enlarged.

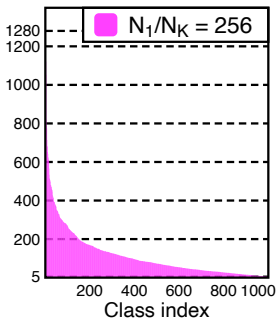


Figure 7. Class distribution of ImageNet-LT.