	CE	Focal[1]	CB[2]	LDAM[3]	BNN[4]	LWS[5]	CE-IC (ours)	CE-DRW-IC (ours)
Resnet50	38.00	38.33	38.88	35.42	33.71	34.10	$\mathbf{32.16 \pm 0.41}$	$\mathbf{32.06 \pm 0.38}$

Table 1: **Validation error↓ on iNaturalist2018.** [1] Focal,Tsung-Yi Lin et al (ICCV 2017); [2] CB, Cui Yin et al (ICCV'2019); [3] LDAM, Kaidi Cao et al (NIPS'2019); [4] BBN, Boyan Zhou et al (CVPR'2020); [5] Decoupling, Kang et al (ICLR' 2020). All methods use resnet 50 and are trained for 90 epochs.

We thank the reviewers for their thorough reviews and positive comments about the novelty, effectiveness and adaptability of the method. We will make corresponding changes to reflect the comments.

Paper Summary **R2:** **Our main contribution (Sec.3.2) is a rebalance method for class imbalance which is a specific (and difficult) form of label prior shift, and is not domain adaptation; label prior shift in the paper refers specifically to change in empirical class frequencies between source/target distribution.** Domain adaption usually dose not consider class imbalance explicitly. We follow the standard datasets and works within this subfield [2][3][4][5], exhibiting ***only*** label prior shift (line 113-117). Specifically, $P_s(Y)$ is the source class priors obtained by counting the number of examples from each class in the training data and $P_t(Y)$ is the class priors for the testing data. The definition of *label prior shift* (line 112-113) and *non-semantic likelihood shift (NSLS)* (line 119-121) do not contradict each other because they constitute different parts of a joint distribution and can occur simultaneously. When both shifts occur, optimality is not guaranteed (which we will mention in the paper) but empirically the performance is still strong (Exp. Sec.4.4.2). **Our second contribution (Sec.3.3) is demonstrating the *adaptability* of the rebalance method by combing it with a *multi-modal fusion algorithm***. Unlike domain adaptation, we don't assume to have unlabeled data in the target domain during training. The fusion algorithm deals with NSLS by *weighting* the modality affected by NSLS whereas domain adaption aims to *adapt* to domain shift utilizing additional target domain data.

Does it also calibrate the network? **R1:** The imbalance calibration technique (Sec.3.2) *rebalances* a network. However, the probabilistic nature means that it can be combined with other probabilistic techniques. One reason of combining with UNO (Sec.3.3) is to demonstrate the **adaptability** of this method since UNO uses the exactly same temperature scaling technique for calibrating a network.

More Comparisons to SOTA **R2 R3:** Our paper proposes a rebalance method (sec.3.2) for class imbalance and compared to those SOTA methods. We thank the reviewers for pointing out works we missed and report them here. We believe that **our method is more effective and easier to adapt than current SOTA methods** as shown in table 1.

Inclusion of non-semantic likelihood shift (NSLS) **R2 R3:** To clarify, Section 3.3 is meant to demonstrate that our rebalance technique (Sec.3.2) is general, and can be *combined* with existing probabilistic methods for mutli-modal fusion e.g, UNO *by considering class imbalance in semantic segmentation*. The mechanism by which the fusion algorithm deals with NSLS is temperature scaling. In a mutli-modal fusion setting, when one modality is under NSLS its prediction is no longer reliable. The fusion algorithm flattens the distribution affected by NSLS and effectively diminishes its contribution when fused with other distributions. It resembles a conventional "gating" mechanism which filters out the degraded modality. We realize that this section requires more background knowledge will add proper introduction and expand explanation in the main paper.

Clarification on Theorem 1 **R2:** The first sentence "Given that $h_s(x)$ is the Bayes classifier ..." is equivalent to "Given $P_s(Y|X)$ is the posterior distribution of the source dataset" because in Eq.2, $h_s(x)$ is defined in terms of $P_s(Y|X)$. While Saerens et al.2002 arrives at the same well-known equation through Bayes Rule as we did, Theorem 1 proved its **optimality** through Bayes Risk. The major contribution of Saerens is an EM algorithm to estimate the unkonwn $P_t(Y)$. **R3**: We agree that if we have the true target distribution, the problem can be solved. However knowing that *we do not have access to the true target distribution*, we propose to find a better approximation to the true distribution. **Intuitively speaking, $P_r(y|x)$ often *over-emphasizes small classes* while $P_d(y|x)$, which fits the imbalanced source data, naturally *biases towards large classes* on test data (line 150-152).** This observation motivates us to balance the two posteriors. Therefore, Hypothesis 1 states that **a better approximation to the true target distribution** can be found by trading off $P_r(y|x)$ and $P_d(y|x)$ through optimization. It is validated empirically through subsequent experiments since varying λ yields better performance.

Clarification on Experimental Setting **R1:** we report our re-trained model performance for other methods using exactly the same code from the authors. The performance is consistent with other papers using the same code. Even considering the original performance, our method still outperforms other methods. **R2:** For iNatrualist and CIFAR datasets, we use the official train/validation splits. For Synthia datasets we split the train/test/validation according to the $7:2:1$ ratio. For CIFAR experiments we report the validation errors as in other papers and use the the same train and validation for all methods. **R2 R3:**Synthia is a semantic segmentation dataset and uses a very different architecture and training schedule compared to image classification datasets. **For fair comparison, we compared to imbalance losses but not methods requiring significant changes to training strategy or architecture (line 268-269) since most imbalance methods are developed for image classification but not semantic segmentation.** This shows that our rebalance method is more ***general*** and applicable to a broader range of problems than current methods.