Towards Calibrated Multi-label Deep Neural Networks

Jiacheng Cheng Nuno Vasconcelos
Department of Electrical and Computer Engineering
University of California, San Diego
{jicheng,nvasconcelos}@ucsd.edu

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.02605

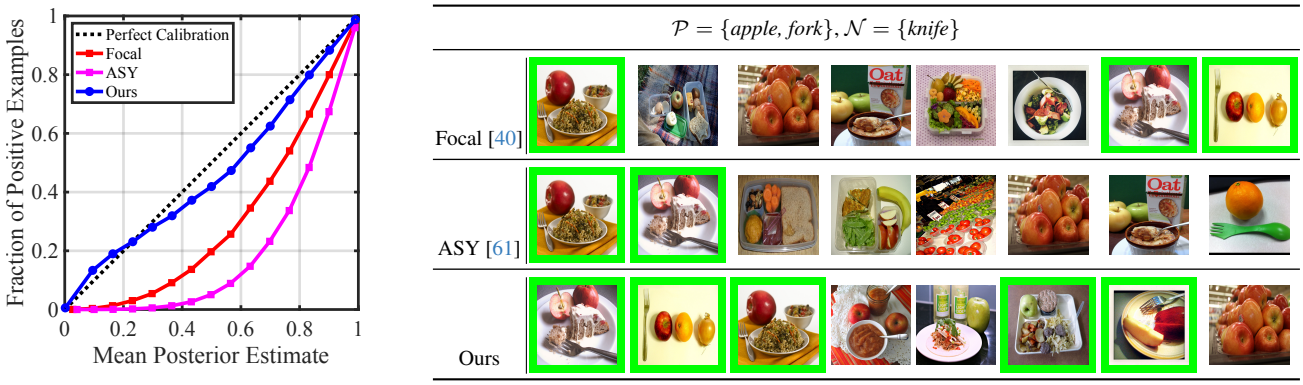

Figure 1. Left: Reliability diagram (calibration curve) of multi-label DNNs trained with the asymmetric focal loss [40], ASY loss [61], and our proposed loss. Right: corresponding retrieval results on the multi-label retrieval task, where the user specifies a query string of desired labels $\mathcal{P}$ and undesired labels $\mathcal{N}$. Correct retrieval results are highlighted in green. Improved calibration substantially improves retrieval performance.

Abstract

The problem of calibrating deep neural networks (DNNs) for multi-label learning is considered. It is well-known that DNNs trained by cross-entropy for single-label, or one-hot, classification are poorly calibrated. Many calibration techniques have been proposed to address the problem. However, little attention has been paid to the calibration of multi-label DNNs. In this literature, the focus has been on improving labeling accuracy in the face of severe dataset unbalance. This is addressed by the introduction of asymmetric losses, which have became very popular. However, these losses do not induce well calibrated classifiers. In this work, we first provide a theoretical explanation for this poor calibration performance, by showing that these loses losses lack the strictly proper property, a necessary condition for accurate probability estimation. To overcome this problem, we propose a new Strictly Proper Asymmetric *(SPA) loss. This is complemented by a* Label Pair Regularizer *(LPR) that increases the number of calibration constraints introduced per training example. The effectiveness of both contributions is validated by extensive experiments on various multi-label datasets. The resulting training method is shown to significantly decrease the calibration error while maintaining state-of-the-art accuracy.*

1. Introduction

Deep neural networks (DNNs) including convolutional neural networks (CNNs) [27, 35] and vision transformers (ViTs) [13] have demonstrated great capacity for solving supervised learning tasks in computer vision. However, many applications require trust-worthy machine learning systems, which are not only accurate but also probability calibrated, *i.e*. able to produce accurate estimates of the posterior probabilities of the various classes. A classifier is calibrated if it predicts a posterior class probability of p when the selection of the class is correct $p \times 100\%$ of the time. The importance of calibration has been noted for many applications. For example, in medical diagnosis [44, 82], probabilities can be used to determine which examples require human inspection, thus avoiding the cost of manually inspecting all images. However, the process can only be trusted if the DNN provides accurate posterior estimates. The safety-critical nature of the application makes probability calibration a critical requirement to enable this functionality. Cost-sensitive applications [15], *e.g*. fraud detection [2, 50] or business decision making [56, 75], involve

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.02605

different costs for different types of misclassification errors. In such cases, accurate class-posterior probabilities are indispensable to building a Bayes optimal decision rule. Therefore, DNN calibration has attracted substantial attention from the computer vision and machine learning community [18, 24, 32, 36, 54, 67].

While various calibration techniques have been proposed, they mostly address DNN training for single-label, or one-hot, classification, where each example has one and only one label. However, various applications require multi-label training. The classical example is image tagging, where a natural image is tagged with a plurality of labels, corresponding to objects or attributes of interest [21]. Other examples include visual question answering (VQA) [22], where a given image/question pair may have multiple correct answers. Multi-label learning has been the subject of extensive research with the focus of improving accuracy, either by designing novel loss [3, 21, 33, 61, 80] or leveraging auxiliary information [8, 12, 25, 79, 83]. Multi-label DNNs can be trained with class probability estimation (CPE) losses that encourage probability calibration, such as the binary cross-entropy (BCE) loss. However, the multi-label setting is highly imbalanced, due to the sparseness of positives, as most tags are absent from any given image. In result, asymmetric losses such as such as the focal loss of [40] or the asymmetric (ASY) loss of [61] tend to produce much higher labeling accuracy than the BCE [61].

While this has made the ASY loss quite popular for multi-label applications such as object detection [23, 48], multi/cross-modal learning [26, 38, 70], or medical diagnosis [5, 31, 46, 84], the question of whether this and similar losses encourage accurate label probability estimates remains unanswered. Besides concerns of trust, safety, or costs, multi-label learning has the challenge of relying on independently predicted tag posterior probabilities. For decisions involving multiple labels, calibration errors can accumulate, degrading the performance of even the most basic operations. We illustrate this with the multi-label retrieval application of Figure 1, where users can retrieve images using a tag-string that implements a conjunction of tag values, *e.g*. "pictures that contain fork and apple, but not knife". These type of queries are of interest for applications like wildlife biology, where a user may seek images containing animals of two or more species to study their interactions, face recognition systems that enable attribute based search, *e.g*. people that "have an elongated chin but do not wear eyeglasses", *etc*. For these applications, calibration becomes an essential requirement, in the sense that it affects even the accuracy of the retrieval operation.

Our preliminary studies reveal that multi-label DNNs trained with existing losses tend to produce poorly-calibrated probabilities. This is illustrated by the calibration curves in Figure 1, where it can be seen that the calibration of the focal and ASY losses are drastically far from perfect. We argue that these popular multi-label losses are poorly suited for class-probability estimation because they are not *strictly proper* [4, 19, 68]. This is a property that denotes the family of losses uniquely minimized by the true posterior probability. It is know in statistical learning that, even in the asymptotic limit of infinite training data, the probability estimates provided by a classifier can only be trusted if the latter is trained with a strictly proper loss [42, 59]. In summary, existing multi-label losses fail to address all requirements needed to support applications like multi-label retrieval. On one hand, classical losses like the BCE are strictly proper but cannot handle the asymmetry of multi-label datasets. On the other, asymmetric losses, such focal or ASY, are robust to this asymmetry but lack the strictly proper property needed to ensure probability calibration.

This motivated us to seek a new multi-label loss that is simultaneously asymmetric and strictly proper. We introduce the *Strictly Proper Asymmetric* (SPA) loss to satisfy these constraints. Extensive experiments demonstrate that it produces significantly better calibrated probability estimates than existing multi-label losses, as illustrated in Figure 1, without sacrificing their multi-label classification accuracy. However, we further acknowledge that, despite its asymptotic guarantees, the strictly proper property is not sufficient to guarantee calibrated probabilities in the finite training data regime. To improve on this, we leverage the structure of the multi-label problem and propose a new regularizer that encourages consistency of the label pair probabilities. This is denoted as the *label pair regularizer* (LPR). This is shown to further improve calibration without classification cost. Finally, it is shown that improved calibration enables significantly better performance in problems like multi-label retrieval, as exemplified in Figure 1.

The contributions of the paper can be summarized as:

- Theoretical analysis of the poor calibration of DNNs trained with popular asymmetric losses, showing the latter are not strictly proper.
- New SPA loss that is both asymmetric and strictly proper, and new LPR regularizer to improve calibration in the finite data regime.
- Use of the multi-label retrieval task as a testing ground for the importance of multi-label probability calibration.
- Extensive empirical evaluations on multiple multi-label datasets, DNN architectures (CNNs and transformers), showing that both SPA and LPR improve calibration without sacrifice of multi-label classification accuracy.

2. Related Works

Probability Calibration of DNNs. It is known that DNNs trained via supervised learning are frequently poorly-calibrated and over-confident [24, 54]. Many calibration techniques have been proposed, including post-hoc process-

ing (*e.g.* temperature scaling [24]), Bayesian DNNs [18, 32, 67, 69], and train-time regularizations [9, 34, 47, 55, 72, 81]. These techniques are designed for the single-label, or one-hot, classification setting but their effectiveness for the multi-label setting remains to be investigated.

Proper Class Probability Estimation (CPE) Loss. The design of proper scoring rules for probability elicitation has long been investigated in statistics [29, 64]. The family of proper CPE losses was later investigated in machine learning [4, 42, 43, 59, 60]. In statistical learning, asymmetric costs or data imbalance are usually addressed by adoption of asymmetric losses and conditional risks or asymmetric inverse link functions [1, 62, 66, 78]. The (symmetric) focal loss [40] is popular in computer vision tasks but has been shown not to be a strictly proper CPE loss [6]. To the best of our knowledge, the proper property has not been studied for the asymmetric focal losses [40, 61] commonly used for multi-label learning.

Multi-label Classification (MLC). MLC is a special case of multidimensional function learning, long studied in the machine learning literature [30, 41, 85]. Classical MLC approaches can be categorized into *transformation-based classifiers* (*e.g.* binary relevance [20], classifier chain [58]) that transform MLC into binary or multiclass problems and *adaptation-based classifiers* (*e.g.* AdaBoost.MH [65], rank-SVM [14]) that adapt popular learning algorithms for MLC, are usually only applicable to shallow models. Most DNN-based MLC work has focused on either designing losses (including both CPE losses [3, 61, 80] and non-CPE losses [21, 33]) or leveraging auxiliary information (*e.g.* class co-occurrence [8, 12, 79, 83], text description [25], spatial annotations [63, 76, 86]) to improve classification accuracy. In this work, we focus on the design of losses that encourage multi-label DNNs to be both accurate and probability calibrated, without the use of auxiliary information.

3. Towards Calibrated Probabilistic Multi-label DNNs

3.1. Preliminaries

Notations. Let $\mathbb{R} = (-\infty, +\infty)$, $\mathbb{R}_+ = [0, +\infty)$, $\mathbb{R}_{++} = (0, +\infty)$, and $\Delta = [0, 1]$. For any $x \in \mathbb{R}$, we denote $\max(x, 0)$ by $(x)_+$. Given event A, the indicator function $\mathbb{1}_A$ has value 1 if A is true and 0 otherwise.

In multi-label classification, each example $\mathbf{x} \in \mathcal{X}$ can have multiple labels simultaneously. The label $\mathbf{y} = [y^{(1)}, \cdots, y^{(T)}]^\top \in \mathcal{Y} = \{-1, +1\}^T$ is a vector of T binary labels or tags. Modern multi-label DNNs usually predict labels separately [10, 61, 70, 73] under the following independence assumption.

Assumption 1. *For any* $\mathbf{x} \in \mathcal{X}$ *and* $i \neq j \in \{1, \cdots, T\}$, $y^{(i)}$ *and* $y^{(j)}$ *are independent given* $\mathbf{x}$, i.e. $y^{(i)} \perp\!\!\!\perp y^{(j)} | \mathbf{x}$.

While this assumption is imperfect and fails to capture label dependence, it underlies the popular multi-label losses discussed in this work and is thus necessary for the theoretical analysis of whether they are proper for probability estimation. For this reason, we adopt this assumption in this work and leave the study of proper losses for joint class probability estimation over multiple labels to future work.

Given observation $\mathbf{x} \in \mathcal{X}$, the goal is to estimate the vector $\boldsymbol{\eta}(\mathbf{x}) = \left[\eta^{(1)}(\mathbf{x}), \cdots, \eta^{(T)}(\mathbf{x})\right]^\top$ of class-posterior probabilities

$$\eta^{(t)}(\mathbf{x}) = P\left(y^{(t)} = 1 | \mathbf{x}\right), \forall t \in \{1, \cdots, T\}. \quad (1)$$

A multi-label DNN typically performs this probability estimation in two steps. First, it maps $\mathbf{x} \in \mathcal{X}$ into a real-valued score vector $\mathbf{v}(\mathbf{x}) = \left[v^{(1)}(\mathbf{x}), \cdots, v^{(T)}(\mathbf{x})\right]^\top \in \mathbb{R}^T$. The embedding $\mathbf{v} : \mathcal{X} \to \mathbb{R}^T$ is composed by a sequence of linear and nonlinear operations. Each $v^{(t)}(\mathbf{x})$ is then mapped into a class-posterior probability estimate with

$$\widehat{\eta}^{(t)}(x) = \widehat{P}\left(y^{(t)} = 1 | \mathbf{x}\right) = [\Psi]^{-1}(v^{(t)}(\mathbf{x})), \quad (2)$$

where $[\Psi]^{-1}(\cdot)$ is an inverse link function. This can be any strictly increasing function $[\Psi]^{-1} : \mathbb{R} \to \Delta$, but is usually the logistic inverse link, or sigmoid activation function

$$\sigma(v) = \frac{1}{1 + e^{(-v)}}, \quad (3)$$

in which case $v^{(t)}(\mathbf{x})$ is called a logit.

Given a CPE loss $\ell : \Delta \times \{-1, +1\} \to \mathbb{R}$ that assigns a cost $\ell(\widehat{\eta}, \pm 1)$ for predicting $\widehat{\eta}$ as the class-posterior probability of positive class when the true label is $y = \pm 1$,[1] the optimal posterior probability estimator minimizes the *risk*:

$$\mathcal{R}(\widehat{\boldsymbol{\eta}}) = \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\sum_{t=1}^{T} \ell\left(\widehat{\eta}^{(t)}(\mathbf{x}), y^{(t)}\right)\right], \quad (4)$$

where $\mathbb{E}$ denotes expectation. To train a multi-label probabilistic DNN, this is approximated by the *empirical risk*

$$\widehat{\mathcal{R}}(\widehat{\boldsymbol{\eta}}; \mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \ell(\widehat{\eta}^{(t)}(\mathbf{x}_i), y_i^{(t)}) \quad (5)$$

on a training dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ of i.i.d. samples from $\mathcal{X} \times \mathcal{Y}$ [74].

A CPE loss can be equivalently expressed with a pair of *partial losses* $\ell_{\pm 1}(\widehat{\eta}) := \ell(\widehat{\eta}, \pm 1)$. Using (2), $\ell_{\pm 1}(\widehat{\eta})$ can be rewritten as the *composite loss* $\ell_{\pm 1,\Psi}(v) := \ell_{\pm 1}([\Psi]^{-1}(v))$, which is a function of a real-valued score $v \in \mathbb{R}$. In this work, we use $\ell_{\pm 1}(\widehat{\eta})$ and $\ell_{\pm 1,\Psi}(v)$ to denote the CPE loss and the CPE composite loss, respectively.

[1] We sometimes omit dependencies on $\mathbf{x}$ and t for notation simplicity.

3.2. Multi-label Learning Loss

The most popular multi-label loss is the binary cross-entropy (BCE)

$$\begin{cases} \ell_{+1}^{BCE}(\widehat{\eta}) = -\log(\widehat{\eta}), \\ \ell_{-1}^{BCE}(\widehat{\eta}) = -\log(1-\widehat{\eta}). \end{cases} \tag{6}$$

This is symmetric, in that $\ell_{+1}^{BCE}(\widehat{\eta}) = \ell_{-1}^{BCE}(1-\widehat{\eta})$. However, real-world multi-label datasets tend to have highly imbalanced label distributions, since negative labels (absence of a tag) are much more frequent than positive ones (presence of tag). This imbalance degrades BCE loss performance in comparison to asymmetric losses, such as the focal loss [40, 61]

$$\begin{cases} \ell_{+1}^{AsyFocal}(\widehat{\eta}) = -(1-\widehat{\eta})^{\gamma^+}\log(\widehat{\eta}), \\ \ell_{-1}^{AsyFocal}(\widehat{\eta}) = -(\widehat{\eta})^{\gamma^-}\log(1-(\widehat{\eta})), \end{cases} \tag{7}$$

where $\gamma^+, \gamma^- \in \mathbb{R}_{++}$ are focusing parameters for positive and negative examples respectively. For $\gamma^+ < \gamma^-$, this loss assigns relatively higher weights to hard negative examples, discounting the large numbers of negatives away from the boundary. This increases the classification margin and improves multi-label classification performance. [61] proposed to further increase negative example margins by augmenting this loss with asymmetric probability shifting, *i.e*. hard thresholding of the margin to fully discard easy negative examples. This leads to the asymmetric (ASY) loss

$$\begin{cases} \ell_{+1}^{ASY}(\widehat{\eta}) = -(1-\widehat{\eta})^{\gamma^+}\log(\widehat{\eta}), \\ \ell_{-1}^{ASY}(\widehat{\eta}) = -(\widehat{\eta}-m)_+{}^{\gamma^-}\log(1-(\widehat{\eta}-m)_+), \end{cases} \tag{8}$$

where $m \in \mathbb{R}_+$ is the probability margin, and the ASY loss reduces to the asymmetric focal loss of (7) if $m = 0$.

The ASY loss achieves state-of-the-art accuracy on multiple multi-label datasets and has become increasingly popular in machine learning and computer vision [5, 23, 26, 31, 38, 46, 48, 70, 84]. However, we observe that the probability estimates it produces are usually highly uncalibrated (as illustrated by the calibration curve in Figure 1). This makes it poorly suited for applications that require calibrated probability estimates.

3.3. Strictly Proper Asymmetric Loss

Since (4) can be rewritten as

$$\mathcal{R}(\widehat{\boldsymbol{\eta}}) = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathbf{y}|\mathbf{x}}\left[\sum_{t=1}^{T}\ell\left(\widehat{\eta}^{(t)}(\mathbf{x}), y^{(t)}\right)\middle|\mathbf{x}\right]\right] \tag{9}$$

$$= \mathbb{E}_{\mathbf{x}}\left[\sum_{t=1}^{T} C\left(\eta^{(t)}(\mathbf{x}), \widehat{\eta}^{(t)}(\mathbf{x})\right)\right] \tag{10}$$

where C is the (pointwise) *conditional risk*

$$C\left(\eta(\mathbf{x}), \widehat{\eta}(\mathbf{x})\right) = \eta(\mathbf{x})\ell_{+1}(\widehat{\eta}(\mathbf{x})) + (1-\eta(\mathbf{x}))\ell_{-1}(\widehat{\eta}(\mathbf{x})). \tag{11}$$

It follows that the risk (4) is minimized when the conditional risk is minimum for all $\mathbf{x} \in \mathcal{X}$. For $\widehat{\eta}$ to be an accurate estimate of the true class-posterior probability η, $C(\eta, \widehat{\eta})$ should be uniquely minimized by $\widehat{\eta} = \eta$ for all $\eta \in \Delta$. When this holds the CPE loss is said to be *strictly proper*.

Definition 1. *(Strict Properness [4, 19, 68]) The pair of partial losses $\{\ell_{-1}, \ell_{+1}\}$ or $\{\ell_{-1,\Psi}, \ell_{+1,\Psi}\}$ is strictly proper if the conditional risk $C(\eta, \widehat{\eta})$ of (11) is uniquely minimized by $\widehat{\eta} = \eta$ for all $\eta \in [0, 1]$.*

It is well-known that the BCE loss of (6) is strictly proper [43, 59]. We next show that both the focal loss of (7) and the ASY loss of (8) are not strictly proper.

Theorem 1. *For any $m \in \mathbb{R}_+$ and $\gamma^+, \gamma^- \in \mathbb{R}_{++}$, the ASY loss of (8) is not strictly proper.*

Corollary 1. *For any $\gamma^+, \gamma^- \in \mathbb{R}_{++}$, the focal loss of (7) is not strictly proper.*

To the best of our knowledge, Theorem 1 is the first analysis of the strict properness property for the asymmetric focal and ASY losses. Note that when $\gamma^+ = \gamma^-$, it follows from Corollary 1 that the (symmetric) focal loss [40] is not strictly proper, a result first proven in [6, Theorem 5]. [6, Theorem 11] shows that, for the (symmetric) focal loss, the lack of strict properness property can be circumvented, by deriving a bijective map between the conditional risk minimizer and the true class-posterior probability. The following theorem generalizes this result, by showing that the class-posterior probability can be recovered from the minimizer of the conditional risk of the ASY loss (8) by a bijective map if and only if the probability margin is $m = 0$.

Theorem 2. *Denote the conditional risk of (11) defined by the ASY loss ℓ^{ASY} of (8) by $C^{ASY}(\eta, \widehat{\eta})$. Let*

$$\widehat{\eta}^{ASY,*}(\mathbf{x}) = \underset{\widehat{\eta}\in[0,1]}{\arg\min}\, C^{ASY}\left(\eta(\mathbf{x}), \widehat{\eta}(\mathbf{x})\right) \tag{12}$$

be the minimizer of this risk for any $\mathbf{x} \in \mathcal{X}$. Then there is a mapping ϕ such that

$$\eta(\mathbf{x}) = \phi\left(\widehat{\eta}^{ASY,*}(\mathbf{x}); \gamma^-, \gamma^+, m\right), \tag{13}$$

where

$$\phi(z; \gamma^-, \gamma^+, m) = \frac{h((z-m)_+; \gamma^-)}{h((z-m)_+; \gamma^-) + h(1-z; \gamma^+)}, \tag{14}$$

$$h(z; \gamma) = \frac{z^\gamma - \gamma z^{\gamma-1}(1-z)\log(1-z)}{1-z}. \tag{15}$$

$\phi : \Delta \to \Delta$ is a bijective map if and only if $m = 0$.

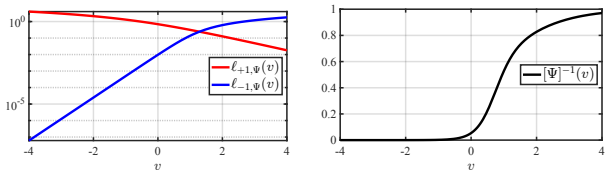

Figure 2. Example of the SPA losses (left) and the associated inverse link (right) where $(\zeta^+, k^+, b^+, \zeta^-, k^-, b^-) = (1, 1, 0, 5, 3, 1)$.

Our empirical studies show that applying the mapping ϕ of (13) significantly improves the calibration of multi-label DNNs trained by both (7) (where $m = 0$) and (8) (where $m > 0$). However, ϕ is bijective and strictly order-preserving (*i.e.* it will preserve the accuracy of the trained DNNs) if and only if $m = 0$. When $m > 0$, the accuracy of multi-label DNNs may be compromised by ϕ. To overcome this limitation, we seek to design a new loss that is simultaneously asymmetric and strictly proper. For this, rather than designing a CPE loss $\ell_{\pm 1}(\widehat{\eta})$ explicitly, we consider the separate design of a composite loss $\ell_{\pm 1,\Psi}(v)$ and an inverse link $[\Psi]^{-1}(v)$ as below

$$\begin{cases} \ell_{+1,\Psi}(v) = -\frac{1}{\zeta^+} \log\left(\frac{1}{1+e^{(-k^+(v-b^+))}}\right), \\ \ell_{-1,\Psi}(v) = -\frac{1}{\zeta^-} \log\left(\frac{1}{1+e^{(+k^-(v-b^-))}}\right), \end{cases} \tag{16}$$

$$[\Psi]^{-1}(v) = \frac{\frac{k^-}{\zeta^-}\sigma(-k^-(v-b^-))}{\frac{k^-}{\zeta^-}\sigma(-k^-(v-b^-)) + \frac{k^+}{\zeta^+}\sigma(k^+(v-b^+))}, \tag{17}$$

where $\sigma(\cdot)$ is the sigmoid function of (3). According to the Theorem 3 below, the CPE loss composed of (16) and (17) is strictly proper and thus referred to as the *Strictly Proper Asymmetric* (SPA) loss in this work. In practice, following the practice in the literature [61], we reduce the positive partial loss to the BCE loss by setting $(\zeta^+, k^+, b^+) = (1, 1, 0)$ and only tune the hyperparmeters ζ^-, k^-, b^- of the negative partial loss. The motivation for these hyperparmeters is simple and intuitive: i) k^-, b^- define an affine transformation of the logits $v^{(t)}(\mathbf{x})$ that enables control of the rate at which ℓ_{-1} decays to 0 as $v \to 0$. ii) ζ^- is a scale factor that controls the overall weight of negative examples. Note that unlike prior losses (6)-(8), SPA does not directly operate on the probability estimate $\widehat{\eta}(\mathbf{x})$, and that the introduction of these hyperparameters induces the need for the inverse link of (17) to achieve the strict properness, rather than simply using the sigmoid of (3).

Theorem 3. *For any $\zeta^+, \zeta^-, k^+, k^- \in \mathbb{R}_{++}$ and $b^-, b^+ \in \mathbb{R}$, the CPE loss composed of composite loss (16) and inverse link function (17) is strictly proper.*

Finally, Figure 2 presents an example of the SPA loss and Table 1 summarises the properties of the CPE losses discussed in this section. The proofs of all theoretical results in this section are included in the Appendix.

	Definition	Asymmetric	Strict Proper	Recover η
BCE	(6)	✗	✓	✓
Focal [40]	(7)	✓	✗	✓
ASY [61]	(8)	✓	✗	✗
SPA	(16,17)	✓	✓	✓

Table 1. Multi-label CPE losses discussed in this work. The last column indicates whether there is a bijective map between the true class-posterior probability η and the CPE risk minimizer.

3.4. Label Pair Regularizer

A strictly proper CPE loss only guarantees perfect label-probability estimates for asymptotically large datasets. In practice, for finite datasets, empirical risk minimization does not guarantee the recovery of true risk minimizer. In this case, probability calibration can usually be improved by adding regularization terms to the loss function. In this work, we propose a regularizer specifically designed for multi-label learning.

Under the independence Assumption 1, the probability estimates of each label in $\{1, \cdots, T\}$ are supervised independently, *i.e.* there is no explicit supervision for the joint prediction of multiple classes. This is consistent with the decomposition of the risk of (4) into a sum of t label-specific risks. However, an example $\mathbf{x}$ still provides joint constraints on the probability estimates of different labels. To see this, consider any $\mathbf{x} \in \mathcal{X}$ and pair (t, t') of labels with different values. The probability of $y^{(t)} = +1$ is then

$$\begin{aligned} \beta^{tt'}(\mathbf{x}) &= P\left(y^{(t)} = +1 | y^{(t)} \neq y^{(t')}, \mathbf{x}\right) \\ &= \frac{P(y^{(t)} = +1, y^{(t')} = -1|\mathbf{x})}{P\left(y^{(t)} = +1, y^{(t')} = -1|\mathbf{x}\right) + P\left(y^{(t)} = -1, y^{(t')} = +1|\mathbf{x}\right)} \\ &= \frac{\eta^{(t)}(\mathbf{x})(1-\eta^{(t')}(\mathbf{x}))}{\eta^{(t)}(\mathbf{x})(1-\eta^{(t')}(\mathbf{x})) + (1-\eta^{(t)}(\mathbf{x}))\eta^{(t')}(\mathbf{x})} \end{aligned} \tag{18}$$

where the last equality follows from the Assumption 1. Let

$$\widehat{\beta}^{tt'} = \frac{\widehat{\eta}^{(t)}(1-\widehat{\eta}^{(t')})}{\widehat{\eta}^{(t)}(1-\widehat{\eta}^{(t')}) + \widehat{\eta}^{(t')}(1-\widehat{\eta}^{(t)})} \tag{19}$$

be the plugin estimator of $\beta^{tt'}$. The following result shows that the accurate estimation of $\widehat{\beta}^{tt'}$ is a necessary condition for the accurate estimation of $\widehat{\eta}^{(t)}$ and $\widehat{\eta}^{(t')}$.

Lemma 1. *For any $t \neq t'$, $\widehat{\beta}^{(tt')} = \beta^{(tt')}$ is a necessary condition for $\widehat{\eta}^{(t)} = \eta^{(t)}$ and $\widehat{\eta}^{(t')} = \eta^{(t')}$.*

Since the example $\mathbf{x}$ can be seen as a calibration constraint for the estimation of $\widehat{\beta}^{tt'}$ in addition to those that it already provides for the individual calibration of $\widehat{\eta}^{(t)}$ and $\widehat{\eta}^{(t')}$, this suggests that introducing calibration supervision on $\widehat{\beta}^{tt'}$ can help improve the calibration of $\eta^{(t)}, \eta^{(t')}$. We leverage this observation by introducing a new *Label Pair Regularizer* (LPR) o calibrate the estimate $\widehat{\beta}^{ij}$, imple-

mented with the BCE loss

$$\mathcal{L}^{\text{LPR}}(\mathbf{x},\mathbf{y}) = \frac{2}{T(T-1)} \sum_{t\neq t'} \mathbb{1}_{\substack{y^{(t)}=+1 \\ y^{(t')}=-1}} \left[-\log \widehat{\beta}^{tt'} \right]. \tag{20}$$

Finally, in our proposed training approach, a multi-label DNN is trained by a joint optimization of SPA and LPR:

$$\mathcal{L}^{\text{overall}}(\mathbf{x},\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} \ell_{y^{(t)},\Psi}(v^{(t)}(\mathbf{x})) + \lambda \mathcal{L}^{\text{LPR}}(\mathbf{x},\mathbf{y}) \tag{21}$$

where λ is the multiplier balancing the two terms.

4. Multi-label Image Retrieval

Multi-label probability calibration has many applications beyond image tagging with binary labels. In fact, binary tagging is not the most demanding application in terms of probability calibration. Many binary classification techniques, *e.g.* support vector machines [74], are known to be accurate for classification but not well calibrated. This has motivated a literature on post-hoc calibration techniques, such as Platt scaling [57]. In general, tasks that involve reasoning about multiple tags benefit more from calibrated probability estimates. Consider the problem of multi-label image retrieval, where a user provides a query string with multiple tags, where a tag can be specified as desired ($y = +1$) or undesired ($y = -1$). For example, the user may want pictures of "dogs" but "not on the beach" and "not on the park". Let the set of positive labels be $\mathcal{P} = \{y^{(j_1)}, \cdots, y^{(j_{|\mathcal{P}|})}\}$ and the set of negative labels $\mathcal{N} = \{y^{(k_1)}, \cdots, y^{(k_{|\mathcal{N}|})}\}$. A natural score for image ranking is then the posterior probability

$$\begin{aligned} s(\mathbf{x}) &= P\left(\{y^{(j)}=+1\}_{j\in\mathcal{P}}, \{y^{(k)}=-1\}_{k\in\mathcal{N}} \middle| \mathbf{x}\right) \\ &= \prod_{j\in\mathcal{P}} \eta^{(j)}(\mathbf{x}) \prod_{k\in\mathcal{N}} (1-\eta^{(k)}(\mathbf{x})) \end{aligned} \tag{22}$$

where the equality follows from the Assumption 1. This can be estimated by a probabilistic multi-label networks as

$$\widehat{s}(\mathbf{x}) = \prod_{j\in\mathcal{P}} \widehat{\eta}^{(j)}(\mathbf{x}) \prod_{k\in\mathcal{N}} (1-\widehat{\eta}^{(k)}(\mathbf{x})). \tag{23}$$

However, since this involves the multiplication of several probability estimates whose errors can accumulate, the error of the estimated score usually increases with the total number $|\mathcal{P}| + |\mathcal{N}|$ of labels included in the query. In practice, the performance of the retrieval operation tends to degrade quickly as the length of the query string increases. Probability calibration can decrease the rate of this decay. We thus use this application to evaluate the practical benefits of both the SPA loss and the LPR regularizer.

	MS-COCO	VOC2012	WIDER-A	VISPR
# classes	80	20	14	68
# train images	82081	5717	28340	14167
# test images	40137	5823	29117	8000
$\frac{\text{\# positive labels}}{\text{\# negative labels}}$	0.04	0.08	0.26	0.08

Table 2. Statistics of multi-label classification datasets used for evaluation.

5. Experiments

In this section, we evaluate the calibration performance of the SPA loss and the LPR calibration regularizer.

5.1. Experiment Setup

Networks and Datasets. We evaluate the benefits of the proposed contributions for both CNNs and transformers, using the ECA-ResNet50-T [27, 28, 77] and ViT-B/32 [13] networks as multi-label DNN backbone architecture, respectively. Four public multi-label image classification datasets are used for the empirical evaluations: MS-COCO [39], PASCAL VOC 2012 [16, 17], WIDER Attribute [37], and MPI-I VISPR [53]. In MS-COCO and VOC, each label indicates the existence of an object class in the image, while in WIDER-A and VISPR, each label indicates the possession of an attribute. More datasets statistics can be found in Table 2. Beyond image classification, we evaluate the proposed contributions on the multi-modal multi-label task of visual question answering on the VQA v2.0 dataset [22] and the LXMERT [71] network. In this case, the DNN is faced with an image/question pair and each label is a possible answer to the question. We compare SPA and the BCE loss, which is used to train most end-to-end VQA DNNs [7, 71], in the Appendix.

Baselines. We use the BCE, (asymmetric) focal, and ASY losses, of Table 1 as baselines in all experiments. These are complemented by the recent two-way loss (TWL) [33], a state-of-the-art non-CPE loss for multi-label learning. For focal and ASY, we also evaluate the composition of the probability estimate with the true class probability recovery mapping ϕ of (12). Since TWL is a non-CPE loss and cannot produce class-posterior probability estimates, we attempt to calibrate its outputs by temperature scaling [24].

Evaluation Metrics. For measuring image classification accuracy, we follow the protocols in the literature [33, 61] and employ two metrics: class-based mean average precision (mAP@$\mathbf{y}$) and example-based mean average precision (mAP@$\mathbf{x}$). The former/latter is obtained by first calculating the average precision (AP) for each class/example and then averaging them, *i.e.* $\text{mAP@}\mathbf{y} = \frac{1}{T}\sum_{t=1}^{T} \text{AP}(\{(v^{(t)}(\mathbf{x}_i), y_i^{(t)})\}_{i=1}^{N})$, $\text{mAP@}\mathbf{x} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}(\{(v^{(t)}(\mathbf{x}_i), y_i^{(t)})\}_{t=1}^{T})$. For evaluating class-posterior probability calibration, we employed two probability calibration metrics based on the reliability diagram [11, 52]: average calibration error (ACE) [51]

Dataset	Method	ECA-ResNet50-T Accuracy mAP@$\mathbf{y}$ ↑	ECA-ResNet50-T Accuracy mAP@$\mathbf{x}$ ↑	ECA-ResNet50-T Calibration ACE ↓	ECA-ResNet50-T Calibration MCE ↓	ViT-B/32 Accuracy mAP@$\mathbf{y}$ ↑	ViT-B/32 Accuracy mAP@$\mathbf{x}$ ↑	ViT-B/32 Calibration ACE ↓	ViT-B/32 Calibration MCE ↓
COCO	BCE	72.2	82.4	8.2	17.0	70.1	81.9	6.0	15.5
	TWL	77.2	87.9	17.2	33.6	76.2	88.0	14.2	31.3
	Focal	74.8	85.9	20.1	34.9	72.0	83.8	23.6	36.2
	Focal+ϕ	-	-	11.1	20.3	-	-	8.5	17.9
	ASY	77.5	88.2	30.6	46.0	76.4	87.7	29.5	48.6
	ASY + ϕ	77.0	87.8	15.0	26.0	76.2	87.3	16.2	26.2
	SPA	**77.8**	88.4	5.3	12.0	**76.8**	87.9	4.8	10.7
	SPA + LPR	77.7	**88.6**	**4.2**	**9.3**	76.6	**88.1**	**2.1**	**5.3**
VOC	BCE	85.1	92.4	6.9	13.8	86.9	92.7	7.8	16.7
	TWL	89.1	93.7	10.8	22.0	90.1	94.5	14.5	28.3
	Focal	87.4	93.3	19.0	35.8	88.4	93.6	16.2	31.7
	Focal + ϕ	-	-	7.0	16.5	-	-	9.3	20.7
	ASY	89.6	**94.6**	26.4	47.7	**90.4**	94.7	31.7	52.8
	ASY + ϕ	89.1	94.3	12.9	24.7	89.8	94.2	15.5	33.0
	SPA	89.5	94.1	5.4	14.9	90.0	93.9	6.1	14.4
	SPA + LPR	**89.9**	94.3	**4.9**	**11.0**	**90.4**	**94.8**	**5.5**	**12.7**
WIDER-A	BCE	74.9	82.8	11.5	25.2	76.7	83.2	8.6	13.3
	TWL	79.9	85.5	14.3	29.0	81.6	87.6	16.7	32.3
	Focal	78.5	84.1	20.1	34.2	80.8	85.2	18.8	31.2
	Focal + ϕ	-	-	9.3	21.0	-	-	7.4	16.5
	ASY	**80.6**	**86.0**	24.2	35.2	82.2	87.8	22.8	36.3
	ASY + ϕ	79.9	85.4	14.3	27.5	81.8	87.2	13.3	20.7
	SPA	80.1	85.8	5.3	11.2	82.0	87.6	4.3	10.5
	SPA + LPR	80.3	85.9	**3.5**	**8.0**	**82.7**	**87.9**	**2.8**	**6.2**
VISPR	BCE	46.3	78.8	9.7	16.6	48.4	80.5	8.2	14.8
	TWL	52.4	84.3	17.9	26.0	**53.6**	85.4	12.4	26.8
	Focal	48.0	81.0	24.0	40.5	50.1	83.2	24.6	35.7
	Focal + ϕ	-	-	8.6	18.2	-	-	9.6	14.1
	ASY	51.6	84.0	28.8	44.2	53.0	85.0	27.7	45.2
	ASY + ϕ	51.4	83.7	14.3	23.7	43.9	52.8	16.4	29.0
	SPA	52.4	84.5	5.8	12.1	53.2	85.3	5.9	10.2
	SPA + LPR	**52.7**	**84.9**	**3.0**	**8.1**	53.4	**85.6**	**2.5**	**7.4**

Table 3. Performance of different losses on different combinations of datasets and networks. For each combination, we highlight the best results in bold and the second best results underlined.

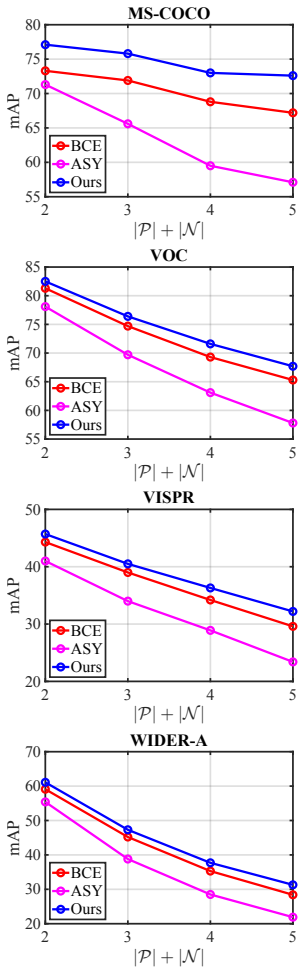

Figure 3. Multi-label image retrieval mAP versus the number of search conditions $|\mathcal{P}| + |\mathcal{N}|$.

and maximum calibration error (MCE) [49] (detailed definition provided in the Appendix). These metrics are suited for evaluating the calibration error under data imbalance [45, 51].

Implementation Details. All multi-label DNNs are trained with a stochastic gradient descent (SGD) optimizer with momentum of 0.9, weight decay of 1e-4, and batch size of 256. The input image resolution is set to 224×224 for both training and testing. Following the protocol of [33, 61], all DNNs are initialized with weights pretrained on ImageNet. For the evaluation of ASY and TWL, we use the official publicly-released implementations and follow the suggestions for hyperparameters choices provided in the original papers. For SPA, we use the hyperparameters $(\zeta^+, k^+, b^+, \zeta^-, k^-, b^-) = (1, 1, 0, 5, 3, 1)$ of Figure 2 and avoid dataset-specific tuning.

5.2. Results

Multi-label Classification. Table 3 summarizes the multi-label classification and calibration performance of all losses in the four multi-label image classification datasets considered. Several observations are possible. First, all losses other than BCE have competitive classification accuracy. The poor performance of BCE is explained by its well known sensitivity to data imbalance. Among the other losses, ASY and SPA tend to have consistently better performance. Second, there are large differences in calibration performance. Compared to the strictly proper losses (SPA and BCE), the TWL, focal, and ASY losses produce significantly worse calibrated models. This confirms our arguments for the importance of the strictly proper property for CPE losses. Third, while the mapping ϕ of (13) significantly improves the calibration of focal and ASY, it is not sufficient to make these losses competitive with BCE and SPA. Finally, SPA has the best calibration among the strictly proper losses, *e.g.* reducing the ACE of the BCE loss by as much as a factor of ≈4 (WIDER-A).

Overall, SPA has accuracy comparable or superior to the current state-of-the-art ASY and TWL with much superior calibration. Regarding LPR, it can be seen that the addi-

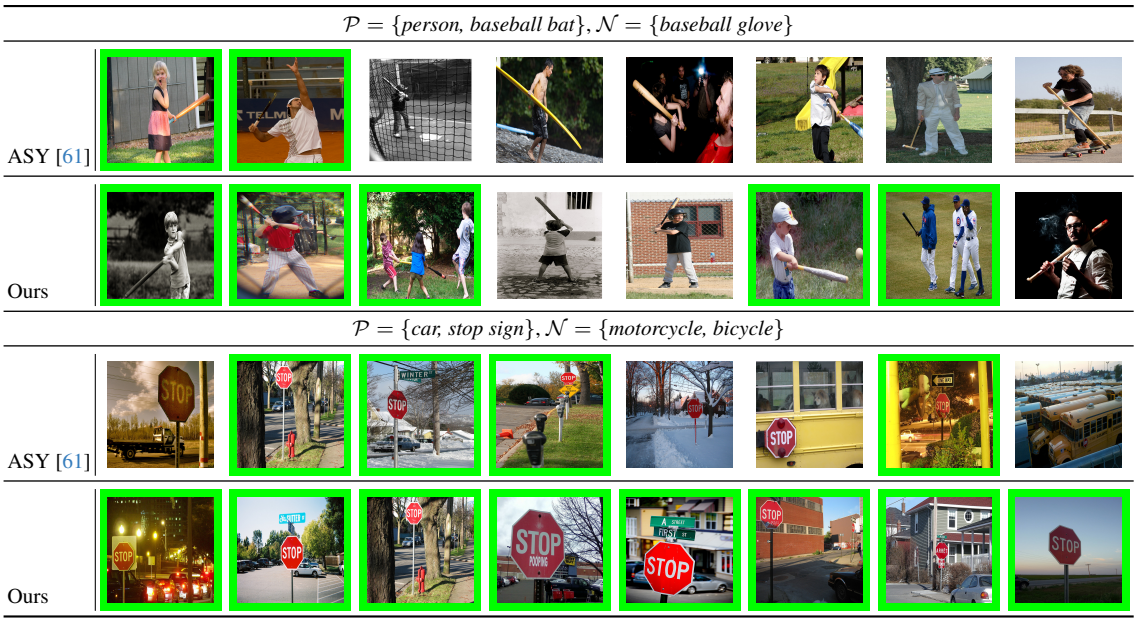

Figure 4. Qualitative results of multi-label image retrieval. Correct retrieval results are highlighted in green.

	mAP@x↑	mAP@y↑	ACE↓	MCE↓
BCE	72.2	82.4	8.2	17.0
BCE + LPR	**72.4**	**82.9**	**6.5**	**14.8**
Focal	74.8	85.9	20.1	**34.9**
Focal + LPR	**75.1**	**86.3**	**19.7**	36.4
ASY	**77.5**	**88.2**	30.6	46.0
ASY + LPR	77.2	88.1	**29.4**	**44.7**

Table 4. Effect of LPR on other CPE losses (COCO, ECA-ResNet50-T).

tion of this regularizer maintains the accuracy of SPA and further enhances calibration, e.g. from 5.8 to 3.0 ACE on VISPER. Altogether, SPA or SPA +LPR achieve the top performance for 29 of the 32 dataset/model/metric configurations of Table 3, while SPA +LPR has top performance for 25 out of the 36. These results demonstrate the effectiveness of SPA +LPR-based multi-label calibration.

Pairwise regularization. While LPR is proposed to complement the SPA loss, it can be used to improve the calibration of any multi-label classification loss. Table 4 summarizes its impact on the prior CPE losses for the ECA-ResNet50-T network on MS-COCO. The addition of the LPR regularizer improves calibration for all losses without substantially altering accuracy.

5.3. Multi-label Image Retrieval

We evaluate the probabilistic multi-label models trained in the last subsection on the multi-label image retrieval task introduced in Section 4. We consider all possible label combinations for $\mathcal{P}$ and $\mathcal{N}$ with $|\mathcal{P}| \geq 1$ and $|\mathcal{P}| + |\mathcal{N}| \leq 5$. The retrieval operation is performed by using (23) to rank all the database images. Performance is evaluated by measuring the mAP over all queries. Figure 3 compares the mAP curves of networks trained with the BCE, ASY, and SPA losses. Note that, even though ASY is a better loss for multi-label classification than the BCE loss, its poor calibration compromises its application to the multi-condition retrieval task, where it severely underperforms the other two losses. The better calibrated multi-label models trained with the SPA loss outperform the networks trained with the other methods in all datasets. The benefits of calibration are also visible in the fact that the performance gap between the SPA networks and the poorly calibrated ASY networks increases with the length of the query string $|\mathcal{P}| + |\mathcal{N}|$. This is consistent with the hypothesis of Section 4. Some qualitative retrieval results are presented in Figure 4.

6. Conclusion

In this work, we investigated the poor calibration of DNNs trained by state-of-the-art multi-label losses, both theoretically and empirically. We explained the poor calibration performance by the lack of the strict properness property for these losses and proposed a new asymmetric loss with this property. This was complemented by a new label pair regularizer that increases the number of calibration constraints per training example. The combination of the two contributions was shown to produce multi-label DNNs featuring both state-of-the-art accuracy and well-calibrated probability estimates.

Acknowledgements

This work was partially funded by NSF award IIS-2303153, a gift from Qualcomm, and NVIDIA GPU donations. We also acknowledge and thank the use of the Nautilus platform for some of the experiments discussed above.

References

[1] Arpit Agarwal, Harikrishna Narasimhan, Shivaram Kalyanakrishnan, and Shivani Agarwal. Gev-canonical regression for accurate binary class probability estimation when one class is rare. In *ICML*, 2014. 3

[2] Alejandro Correa Bahnsen, Aleksandar Stojanovic, Djamila Aouada, and Björn Ottersten. Cost sensitive credit card fraud detection using bayes minimum risk. In *International Conference on Machine Learning and Applications*. IEEE, 2013. 1

[3] Emanuel Ben-Baruch, Tal Ridnik, Itamar Friedman, Avi Ben-Cohen, Nadav Zamir, Asaf Noy, and Lihi Zelnik-Manor. Multi-label classification with partial annotations using class-aware selective loss. In *CVPR*, 2022. 2, 3

[4] Andreas Buja, Werner Stuetzle, and Yi Shen. Loss functions for binary class probability estimation and classification: Structure and applications. Technical report, 2005. 2, 3, 4

[5] Wenjie Cai, Fanli Liu, Bolin Xu, Xuan Wang, Shuaicong Hu, and Mingjie Wang. Classification of multi-lead ecg with deep residual convolutional neural networks. *Physiological Measurement*, 43(7):074003, 2022. 2, 4

[6] Nontawat Charoenphakdee, Jayakorn Vongkulbhisal, Nuttapong Chairatanakul, and Masashi Sugiyama. On focal loss for class-posterior probability estimation: A theoretical perspective. In *CVPR*, 2021. 3, 4

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*. Springer, 2020. 6

[8] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *CVPR*, 2019. 2, 3

[9] Jiacheng Cheng and Nuno Vasconcelos. Calibrating deep neural networks by pairwise constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13709–13718, 2022. 3

[10] Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *CVPR*, 2021. 3

[11] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22, 1983. 6

[12] Zixuan Ding, Ao Wang, Hui Chen, Qiang Zhang, Pengzhang Liu, Yongjun Bao, Weipeng Yan, and Jungong Han. Exploring structured semantic prior for multi label recognition with incomplete labels. In *CVPR*, 2023. 2, 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1, 6

[14] André Elisseeff and Jason Weston. A kernel method for multi-labelled classification. *NeurIPS*, 2001. 3

[15] Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI*, 2001. 1

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, 2010. 6

[17] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015. 6

[18] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 2, 3

[19] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007. 2, 4

[20] Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In *Advances in Knowledge Discovery and Data Mining*, pages 22–30. Springer, 2004. 3

[21] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. In *ICLR*, 2014. 2, 3

[22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017. 2, 6

[23] Inske Groenen, Stevan Rudinac, and Marcel Worring. Panorams: automatic annotation for detecting objects in urban context. *IEEE Transactions on Multimedia*, 2023. 2, 4

[24] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 2, 3, 6

[25] Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Texts as images in prompt tuning for multi-label image recognition. In *CVPR*, 2023. 2, 3

[26] Rohit Gupta, Anirban Roy, Claire Christensen, Sujeong Kim, Sarah Gerard, Madeline Cincebeaux, Ajay Divakaran, Todd Grindal, and Mubarak Shah. Class prototypes based contrastive learning for classifying multi-label and fine-grained educational videos. In *CVPR*, 2023. 2, 4

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 6

[28] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. 2018. 6

[29] Arlo D Hendrickson and Robert J Buehler. Proper scores for probability forecasters. *The Annals of Mathematical Statistics*, 42(6):1916–1921, 1971. 3

[30] Francisco Herrera, Francisco Charte, Antonio J Rivera, María J Del Jesus, Francisco Herrera, Francisco Charte, Antonio J Rivera, and María J del Jesus. *Multilabel classification*. Springer, 2016. 3

[31] Hyeonji Hwang, Seongjun Yang, Daeyoung Kim, Radhika Dua, Jong-Yeup Kim, Eunho Yang, and Edward Choi. Towards the practical utility of federated learning in the medical domain. In *Conference on Health, Inference, and Learning*, pages 163–181. PMLR, 2023. 2, 4

[32] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017. 2, 3

[33] Takumi Kobayashi. Two-way multi-label loss. In *CVPR*, 2023. 2, 3, 6, 7

[34] Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. In *NeurIPS*, 2020. 3

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 1

[36] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. 2

[37] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *ECCV*, 2016. 6

[38] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17990–17999, 2022. 2, 4

[39] 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 *ECCV*, 2014. 6

[40] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 1, 2, 3, 4, 5

[41] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(11):7955–7974, 2021. 3

[42] Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. *NeurIPS*, 2008. 2, 3

[43] Hamed Masnadi-Shirazi and Nuno Vasconcelos. A view of margin losses as regularizers of probability estimates. *Journal of Machine Learning Research (JMLR)*, 16(1):2751–2795, 2015. 3, 4

[44] Alireza Mehrtash, William M Wells, Clare M Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE transactions on medical imaging*, 39(12):3868–3878, 2020. 1

[45] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. 2021. 7

[46] Philip Müller, Felix Meissen, Johannes Brandt, Georgios Kaissis, and Daniel Rueckert. Anatomy-driven pathology detection on chest x-rays. In *MICCAI*, 2023. 2, 4

[47] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *NeurIPS*, 2019. 3

[48] Lorenzo Mur-Labadia, Jose J Guerrero, and Ruben Martinez-Cantin. Multi-label affordance mapping from egocentric vision. In *ICCV*, 2023. 2, 4

[49] Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, 2015. 7

[50] Sanaz Nami and Mehdi Shajari. Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors. *Expert Systems with Applications*, 110: 381–392, 2018. 1

[51] Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Relaxed softmax: Efficient confidence auto-calibration for safe pedestrian detection. In *NeurIPS Workshops*, 2018. 6, 7

[52] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *ICML*, 2005. 6

[53] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *ICCV*, 2017. 6

[54] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019. 2

[55] Hyekang Park, Jongyoun Noh, Youngmin Oh, Donghyeon Baek, and Bumsub Ham. Acls: Adaptive and conditional label smoothing for network calibration. In *ICCV*, 2023. 3

[56] George Petrides, Darie Moldovan, Lize Coenen, Tias Guns, and Wouter Verbeke. Cost-sensitive learning for profit-driven credit scoring. *Journal of the Operational Research Society*, 73(2):338–350, 2022. 1

[57] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999. 6

[58] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85:333–359, 2011. 3

[59] Mark D Reid and Robert C Williamson. Composite binary losses. *Journal of machine learning research (JMLR)*, 11: 2387–2422, 2010. 2, 3, 4

[60] Mark D Reid and Robert C Williamson. Convexity of proper composite binary losses. In *AISTATS*, 2010. 3

[61] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[62] Raúl Santos-Rodríguez, Alicia Guerrero-Curieses, Rocío Alaiz-Rodríguez, and Jesús Cid-Sueiro. Cost-sensitive learning based on bregman divergences. *Machine Learning*, 76: 271–285, 2009. 3

[63] Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. Deep imbalanced attribute classification using visual attention aggregation. In *ECCV*, 2018. 3

[64] Leonard J Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971. 3

[65] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999. 3

[66] Clayton Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6:958–992, 2012. 3

[67] Seonguk Seo, Paul Hongsuck Seo, and Bohyung Han. Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. In *CVPR*, 2019. 2, 3

[68] Emir Shuford, Arthur Albert, and H. Edward Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966. 2, 4

[69] Mahesh Subedar, Ranganath Krishnan, Paulo Lopez Meyer, Omesh Tickoo, and Jonathan Huang. Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference. In *CVPR*, 2019. 3

[70] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. 2022. 2, 3, 4

[71] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, 2019. 6

[72] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *NeurIPS*, 2019. 3

[73] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007. 3

[74] Vladimir Vapnik. *The nature of statistical learning theory*. Springer, 1999. 3, 6

[75] Thomas Verbraken, Wouter Verbeke, and Bart Baesens. A novel profit maximizing metric for measuring classification performance of customer churn prediction models. *IEEE transactions on knowledge and data engineering*, 25(5):961–973, 2012. 1

[76] Meng Wang, Changzhi Luo, Richang Hong, Jinhui Tang, and Jiashi Feng. Beyond object proposals: Random crop pooling for multi-label image recognition. *IEEE Transactions on Image Processing*, 25(12):5678–5688, 2016. 3

[77] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *CVPR*, 2020. 6

[78] Xia Wang and Dipak K. Dey. Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *The Annals of Applied Statistics*, 4(4):2000 – 2023, 2010. 3

[79] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *ICCV*, 2017. 2, 3

[80] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *ECCV*. Springer, 2020. 2, 3

[81] Chen Xing, Sercan Arik, Zizhao Zhang, and Tomas Pfister. Distance-based learning from errors for confidence calibration. In *ICLR*, 2020. 3

[82] Wenjing Yang, Zhantao Cao, Qin Chen, Yuhong Yang, and Guowu Yang. Confidence calibration on multiclass classification in medical imaging. In *ICDM*. IEEE, 2020. 1

[83] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *CVPR*, 2020. 2, 3

[84] Libo Zhang, Lutao Jiang, Ruyi Ji, and Heng Fan. Pidray: A large-scale x-ray benchmark for real-world prohibited item detection. *International Journal of Computer Vision (IJCV)*, pages 1–23, 2023. 2, 4

[85] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 26(8):1819–1837, 2013. 3

[86] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *CVPR*, 2017. 3