Cost-Sensitive Conformal Training with Provably Controllable Learning Bounds

Xuesong Jia[*1], **Yuanjie Shi**[*1], **Ziquan Liu**[2], **Yi Xu**[3], **Yan Yan**[1]

[1]School of EECS, Washington State University
[2]School of EECS, Queen Mary, University of London
[3]Dalian University of Technology
{xuesong.jia, yuanjie.shi, yan.yan1}@wsu.edu, ziquanliu.cs@gmail.com, yxu@dlut.edu.cn

Abstract

Conformal prediction (CP) is a general framework to quantify the predictive uncertainty of machine learning models that uses a set prediction to include the true label with a valid probability. To align the uncertainty measured by CP, conformal training methods minimize the size of the prediction sets. A typical way is to use a surrogate indicator function, usually Sigmoid or Gaussian error function. However, these surrogate functions do not have a uniform error bound to the indicator function, leading to uncontrollable learning bounds. In this paper, we propose a simple cost-sensitive conformal training algorithm that does not rely on the indicator approximation mechanism. Specifically, we theoretically show that minimizing the expected size of prediction sets is upper bounded by the expected rank of true labels. To this end, we develop a rank weighting strategy that assigns the weight using the rank of true label on each data sample. Our analysis provably demonstrates the tightness between the proposed weighted objective and the expected size of conformal prediction sets. Extensive experiments verify the validity of our theoretical insights, and superior empirical performance over other conformal training in terms of predictive efficiency with 21.38% reduction for average prediction set size.

Code — https://github.com/JoSaitama/RWCE

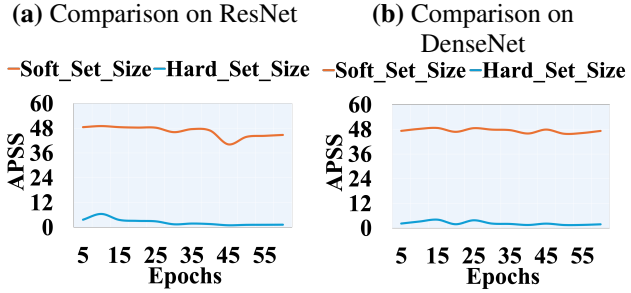

Figure 1: Large approximation error of the learning objective (soft set size) from the true objective (hard set size) during training of ConfTr (Stutz et al. 2021) on CIFAR-100 using ResNet (a) and DenseNet (b). The learning objective is the differentiable measure of average prediction set sizes (APSS) by smooth approximations of the indicator function (e.g., Sigmoid), while the hard set size reflects actual true objectives based on the indicator function. The large and persistent gap between the two highlights the loose approximation introduced by approximated functions, supporting our motivation to avoid indicator approximations and directly optimize a tight upper bound on the true prediction set size.

1 Introduction

Uncertainty Quantification (UQ) (Abdar et al. 2021) is critical for the safe and reliable deployment of modern deep learning models, particularly in high-stakes domains such as medical diagnosis (Begoli, Bhattacharya, and Kusnezov 2019; Yang and Fevens 2021) and autonomous driving (Shafaei et al. 2018; Bachute and Subhedar 2021). In these safety-critical settings, unreliable or missing uncertainty estimates can lead to decisions made without calibrated confidence, increasing the risk of costly errors or harmful outcomes. UQ methods aim to address this challenge by providing confidence-aware predictions that quantify uncertainty alongside predictions. Among UQ methods, Conformal Prediction (CP) (Vovk, Gammerman, and Shafer 2005; Shafer and Vovk 2008; Angelopoulos and Bates 2021) has emerged as a principled, distribution-free framework for constructing reliable prediction sets with formal statistical coverage guarantees.

CP is a powerful and general framework for UQ that transforms the output of a machine learning (ML) model into a prediction set with a user-specified coverage guarantee, typically $1 - \alpha$, without requiring assumptions on the data distribution or model class (Vovk, Gammerman, and Shafer 2005; Angelopoulos and Bates 2021). By leveraging a held-out calibration set, CP ensures that the generated prediction sets contain the true label with high probability, even in finite-sample regimes (Romano, Sesia, and Candes 2020; Angelopoulos et al. 2020a). Owing to its broad applicability and strong theoretical guarantees, CP has found increasing use in critical applications such as fairness-aware learning and robust decision-making (Lu et al. 2022; Lu, Angelopoulos, and Pomerantz 2022). However, a key limitation of standard CP methods is their reliance on post-hoc calibration where the training objective of classification model

[*]These authors contributed equally.

is not minimizing prediction set size, which often results in overly conservative (i.e., large) prediction sets that limit practical utility (Babbar, Bhatt, and Weller 2022; Straitouri et al. 2023).

Recent work (Stutz et al. 2021; Einbinder et al. 2022; Shi et al. 2025; Kiyani, Pappas, and Hassani 2024) directly incorporates the *conformal principles*, e.g., minimizing the size of prediction sets, into model training to produce tighter and more informative prediction sets while preserving valid coverage. Rather than treating calibration as a separate post-hoc step, these methods embed conformal objectives into the learning process (which we refer to as *conformal training*), enabling the model to learn representations that are inherently aligned with uncertainty quantification goals.

One common approach of conformal training is to replace the discrete prediction set size with a differentiable surrogate. It is often referred to as the soft set size and constructed by smooth approximations of the indicator function, e.g., the Sigmoid (Stutz et al. 2021) or Gaussian error function (Kiyani, Pappas, and Hassani 2024). While these relaxations enable gradient-based optimization, they introduce an inherent mismatch: these continuous surrogates cannot precisely approximate the indicator, and do not provide a uniform approximation error bound. As a result, the training loss may deviate significantly from the true prediction set size, leading to suboptimal or unstable behavior during optimization. Figure 1 illustrates this effect on CIFAR-100 with both ResNet and DenseNet backbones: the optimized soft set size (learning objective) remains dramatically separated from the actual hard set size (true objective) throughout training. This raises a natural question: *how could we develop a conformal training method that directly optimizes a tight, theoretically controlled upper bound on the true prediction-set size and dispense with loose surrogates altogether?*

In this paper, we propose a simple yet theoretically grounded cost-sensitive conformal training algorithm that eliminates the need for surrogate approximations of the prediction set indicator. Instead of relying on smooth relaxations, we design a differentiable training objective that leverages the true-label rank to align more closely with the goal of conformal prediction. Our key insight is that the expected size of conformal prediction sets can be tightly upper bounded by the expected rank of the true label under the predictive scores of model. Based on this, we formulate a rank-weighted cross-entropy (RWCE) loss, where each sample is reweighted according to its label rank, yielding a simple rank-weighting scheme that directly incentivizes tighter prediction sets. This objective is efficient to compute, easy to implement with standard gradient-based optimization, and requires no additional smoothing parameters. We theoretically prove that our proposed loss upper bounds both the expected label rank and, consequently, the expected prediction set size, under mild and empirically verifiable conditions. Furthermore, we establish a generalization bound that ensures the empirical loss remains close to its population counterpart with high probability. Experiments on multiple benchmark datasets demonstrate that our method produces significantly smaller prediction sets with an average 21.38% reduction while preserving valid coverage.

Contributions. The key contributions of this paper include:

- We propose a simple yet effective cost-sensitive conformal training algorithm that avoids surrogate indicator approximations. Our method minimizes a RWCE loss, where the importance weight is derived from the true-label rank, providing a differentiable and practical objective for reducing prediction set size.
- We theoretically show that the expected prediction set size is upper bounded by the expected label rank, and that our objective tightly upper bounds this rank under mild conditions. We also establish generalization bounds for RWCE loss, ensuring reliable performance in finite-sample settings.
- Experiments on multiple benchmarks demonstrate that our method consistently produces smaller prediction sets than existing conformal training baselines with an average 21.38% reduction, while maintaining valid coverage guarantees.

2 Related Work

Conformal Prediction (CP). CP (Angelopoulos and Bates 2021; Vovk, Gammerman, and Shafer 2005; Shafer and Vovk 2008) is a distribution-free, post-hoc framework for uncertainty quantification that constructs prediction sets with user-defined coverage guarantees by calibrating conformity scores, which quantify how typical a new instance is relative to a labeled reference set. This framework provides rigorous, finite-sample coverage guarantees. CP can be successfully applied to a wide range of tasks, including regression (Romano, Patterson, and Candes 2019; Gibbs, Cherian, and Candès 2023; Gibbs and Candes 2021), classification (Romano, Sesia, and Candes 2020), computer vision (Bates et al. 2021), federated learning (Lu et al. 2023; Humbert et al. 2023; Plassier et al. 2023, 2024), online learning (Gibbs and Candes 2021; Xi et al. 2025), and adversarial perturbation (Bao et al. 2025; Ghosh et al. 2023). Its performance is primarily evaluated by two critical and often competing criteria: valid coverage and predictive efficiency (Angelopoulos et al. 2021). Recently, research in CP has progressed along two major directions: (i) designing novel nonconformity scores and calibration procedures (Angelopoulos et al. 2021; Huang et al. 2024; Fisch et al. 2021b,a; Guan 2023; Ding et al. 2024; Kiyani, Pappas, and Hassani 2024; Shi et al. 2024; Zhu et al. 2025; Luo and Zhou 2025), and (ii) developing conformal training techniques (Bellotti 2021; Stutz et al. 2021; Einbinder et al. 2022; Yan, Romano, and Weng 2024; Shi et al. 2025), which embed CP principles directly into the learning process to jointly optimize coverage and prediction set size.

Conformal training. Conformal training, initially proposed by Stutz et al. (Stutz et al. 2021), is an emerging direction that aims to integrate the statistical guarantees of conformal prediction (CP) directly into the model training phase. Unlike standard CP approaches that apply calibration as a separate post-hoc procedure, conformal training incorporates conformity-based objectives into the training loop itself (Bellotti 2021; Stutz et al. 2021; Einbinder et al. 2022;

Yan, Romano, and Weng 2024). A central challenge in conformal training lies in formulating a differentiable surrogate for the prediction set size, which is inherently discrete and non-differentiable. To address this, existing approaches such as ConfTr (Stutz et al. 2021), CPL (Kiyani, Pappas, and Hassani 2024), and DPSM (Shi et al. 2025) adopt soft approximations (e.g., Sigmoid or Gaussian error functions) to relax the indicator function that defines whether a label is included in the prediction set. These surrogates allow the use of gradient-based optimization but come at the cost of introducing significant approximation error. As a result, the surrogate objective may poorly reflect the true prediction set behavior, leading to conservative prediction sets, limiting the practical benefits of conformal training. Consequently, despite the promise of integrating CP into end-to-end learning, current surrogate-based approaches struggle to faithfully minimize the quantity that matters most: the size of the conformal prediction set under the coverage constraint. These limitations motivate the need for more principled objectives that better capture the structure of conformal prediction without relying on soft approximations.

3 Problem Setup and Motivation

Notations. Let $X \in \mathcal{X}$ denote an input instance and $Y \in \mathcal{Y} = \{1, 2, \ldots, K\}$ represent its corresponding true class label, where K is the total number of classes. Assume that each pair (X, Y) is drawn i.i.d. from a joint distribution $\mathcal{P}$ over the space $\mathcal{X} \times \mathcal{Y}$. The training dataset is denoted as $\mathcal{D}_{\text{train}} = \{(X_i, Y_i)\}_{i=1}^n$ consisting of n i.i.d. samples. Additionally, we further assume access to two disjoint datasets: a calibration set $\mathcal{D}_{\text{cal}}$ and a test set $\mathcal{D}_{\text{test}}$, where $|\mathcal{D}_{\text{cal}}| = m$. We use $\mathbf{1}[\cdot]$ to denote the indicator function. Additionally, we define a mini-batch $\mathcal{B} = \{(X_i, Y_i) \mid i \in I_s\}$ of size s by sampling indices $I_s \subset \{1, 2, \ldots, n\}$ with $|I_s| = s$. We consider a probabilistic classifier $f \in \mathcal{F}$ that maps inputs to the probability simplex, i.e., $f(X) : \mathcal{X} \to \Delta_+^K$, where Δ_+^K denotes the $(K-1)$-dimensional simplex. The output $f(X)_y$ indicates the predicted confidence (e.g., Softmax score) assigned to class y. Here, $\mathcal{F}$ refers to the hypothesis class from which the classifier is selected. The label rank of y on X predicted by the ML model f is defined by $R_f(X, y) = \sum_{l=1}^K \mathbb{1}[f(X)_l \geq f(X)_y]$. We omit the subscript f when unambiguous.

Conformal Prediction. Conformal prediction (CP) adjusts the outputs of a pre-trained model using a *nonconformity* scoring function. We define a nonconformity score function as $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which quantifies how typical a candidate label is for a given input (Vovk, Gammerman, and Shafer 2005). For any training example $(X_i, Y_i) \in \mathcal{D}_{\text{cal}}$, we write its nonconformity score as $S_{f,i} = S_f(X_i, Y_i)$. Let $S_{f,(j)}$ denote the j-th order statistic (i.e., the j-th smallest value) in the set $\{S_{f,i}\}_{i=1}^m$. A variety of scoring functions have been introduced in the literature (Shafer and Vovk 2008; Angelopoulos et al. 2021; Huang et al. 2024). In this work, we focus on several representative approaches: Homogeneous Prediction Sets (HPS) (Sadinle, Lei, and Wasserman 2019), Adaptive Prediction Sets (APS) (Romano, Sesia, and Candes 2020), Regularized Adaptive Prediction Sets (RAPS) (Angelopoulos et al. 2021), and Sorted Adaptive Prediction Sets (SAPS) (Huang et al. 2024). We assume that the nonconformity scores are distinct (i.e., no ties), following (Romano, Sesia, and Candes 2020).

Given a predefined miscoverage rate $\alpha \in (0, 1)$ and a test input-label pair $(X_{\text{test}}, Y_{\text{test}})$, the goal of CP is to construct a prediction set $\mathcal{C}_f : \mathcal{X} \to 2^{\mathcal{Y}}$ such that the true label is contained in the set with high probability:

$$\mathbb{P}_{(X_{\text{test}}, Y_{\text{test}}) \backsim \mathcal{P}}\{Y_{\text{test}} \in \mathcal{C}_f(X_{\text{test}})\} \geq 1 - \alpha. \tag{1}$$

To achieve this guarantee, the prediction set is formed by thresholding nonconformity scores using a quantile statistic estimated from a held-out calibration dataset $\mathcal{D}_{\text{cal}}$:

$$\widehat{C}_f(X) = \{y \in \mathcal{Y} : S_f(X, y) \leq \widehat{Q}\}$$

where $\widehat{Q}$ is determined by the empirical quantile computed as the $\lceil (1-\alpha)(m+1) \rceil$-th smallest value among $\{S_{f,i}\}_{i=1}^m$, with $m = |\mathcal{D}_{\text{cal}}|$.

Conformal training. Conformal training embeds the desiderata of conformal prediction directly into the learning objective of a classifier f, so that the model is optimized not only for point-wise accuracy but also for the quality of the prediction sets it will output (Bellotti 2021; Stutz et al. 2021; Einbinder et al. 2022; Shi et al. 2025). Among several proposed conformal losses (see Appendix C), we focus on the particularly natural objective that penalizes the expected size of the conformal prediction set:

$$\mathcal{L}_c(f) = \mathbb{E}_X[|\mathcal{C}_f(X)|] = \mathbb{E}_X\Big[\sum_{y \in \mathcal{Y}} \mathbb{1}\big[S_f(X, y) \leq Q\big]\Big]. \tag{2}$$

As the indicator in (2) is non-differentiable, existing methods, such as ConfTr (Stutz et al. 2021) and CPL (Kiyani, Pappas, and Hassani 2024), replace it with a smooth surrogate, e.g., a Sigmoid function. Then the effective conformal loss $\widetilde{\mathcal{L}}_c$ is built by

$$\widetilde{\mathcal{L}}_c(f) = \mathbb{E}_X\Big[\sum_{y \in \mathcal{Y}} \tilde{\mathbb{1}}\big[S_f(X, y) \leq Q\big]\Big], \tag{3}$$

where $\tilde{\mathbb{1}}[\cdot]$ is a smoothed estimator for the indicator function $\mathbb{1}[\cdot]$ and defined by a Sigmoid function (Stutz et al. 2021), i.e., $\tilde{\mathbb{1}}[S \leq q] = 1/(1 + \exp(-(q - S)/\tau))$, or Gaussian error function (Kiyani, Pappas, and Hassani 2024), i.e., $\tilde{\mathbb{1}}[S \leq q] = \frac{1}{2}\big(1 + \text{erf}((q - S)/\sqrt{2}\sigma)\big)$, where $\text{erf}(a) = 2/\sqrt{\pi} \int_0^a e^{-t^2}$ and σ controls the smoothness.

Although smooth relaxations make the prediction set size differentiable, they fail to capture the binary decision structure inherent to CP. Specifically, prediction sets are defined by a hard thresholding rule: a label is either included or excluded based on whether its conformity score falls below a calibrated quantile. Smooth surrogates, such as the Sigmoid or Gaussian error function, can only approximate this decision boundary continuously, and inevitably introduce approximation error near the threshold. Since these surrogates lack uniform approximation guarantees, the training objective may diverge from the true prediction set size, resulting in misaligned gradients and suboptimal performance. Recall that Figure 1, this misalignment manifests empirically

on CIFAR-100 with both ResNet and DenseNet backbones: the learning objective systematically overestimates the true objective across training epochs, resulting in a persistent optimization gap and an unreliable training signal.

Therefore, our goal is to propose a conformal training method which directly optimize a tight, differentiable, and controllable bound on the true prediction set size.

4 Cost-Sensitive Conformal Training

In this section, we introduce our proposed framework, *Rank Weighted Cross-Entropy (RWCE)*, which incorporates label-rank information into the training objective to align with conformal prediction goals. We then provide a theoretical analysis of RWCE, establishing learning bounds that justify its ability to tightly control prediction set size while ensuring reliable generalization.

4.1 Importance Weighting with Label Rank

Our approach avoids loose indicator approximations by directly leveraging the rank of the true label, a discrete statistic that reflects how prominently the correct class is scored by the model. Intuitively, a model that consistently ranks the true label near the top is more likely to produce smaller prediction sets after calibration, as fewer incorrect classes will score above the threshold. Therefore, the expected rank naturally serves as a proxy for prediction set size. The key insight of our work is to formalize and exploit this connection: by minimizing a rank-weighted cross-entropy loss, our method provides a differentiable and practical objective that aligns training directly with the efficiency goal of conformal prediction without relying on any surrogate relaxation.

To incorporate this idea into a differentiable training loss, we introduce an importance weighting scheme via the true-label rank. We rewrite the true label rank as $R_f(X, Y)$ to explicitly show the impact of model f, where it denotes the position of the true label Y in the sorted list of class scores. Crucially, $R_f(X, Y)$ is treated as a fixed, non-differentiable weight: it serves purely as a cost signal, and we do not back-propagate through the rank itself. Harder examples (those ranked higher) incur a larger penalty, while easier examples (those ranked lower) are rewarded. By scaling the standard cross-entropy loss with the rank, we obtain a cost-sensitive loss function that emphasizes reducing the frequency and magnitude of low-rank predictions. This weighting acts as a soft prioritization mechanism: samples with high confidence in the correct class are minimally penalized, while uncertain samples contribute more to the overall loss. Compared to other soft-set-size surrogates that apply uniform smoothing to all samples, our approach focuses optimization effort on correcting misranked examples in a way that is tightly coupled to conformal prediction efficiency. We define the *population loss* for the rank-weighted cross-entropy (RWCE) as:

$$\mathcal{L}(f) \triangleq \mathbb{E}_{(X,Y)\sim\mathcal{D}}\left[R_f(X, Y) \cdot \ell_f(X, Y)\right], \tag{4}$$

where $\ell_f(X, Y)$ is the standard cross-entropy loss.

The corresponding empirical objective is:

$$\widehat{\mathcal{L}}(f) \triangleq \frac{1}{n}\sum_{i=1}^{n} R_f(X_i, Y_i) \cdot \ell_f(X_i, Y_i). \tag{5}$$

Algorithm 1: Conformal Training with Label Rank

1: **Input**: Training dataset $\mathcal{D}_{\text{tr}}$, learning-rate $\eta > 0$, batch size s
2: Randomly initialize the deep neural network f_0
3: **for** $t \leftarrow 0 : T-1$ **do**
4: Randomly sample a batch $\mathcal{B}_t \subset \mathcal{D}_{\text{tr}}$
5: Compute softmax scores $f_t(X_j)$ for all $X_j \in \mathcal{B}_t$
6: Compute true label ranks $R_{f_t}(X_j, Y_j)$ for each (X_j, Y_j)
7: Compute cross-entropy loss $\ell_{f_t}(X_j, Y_j)$ for each (X_j, Y_j)
8: Compute RWCE loss: $\widehat{\mathcal{L}}(f_t) \leftarrow \frac{1}{s}\sum_{j=1}^{s} R_{f_t}(X_j, Y_j) \cdot \ell_{f_t}(X_j, Y_j)$ on batch $\mathcal{B}_t$
9: Update model: $f_{t+1} \leftarrow f_t - \eta \cdot \nabla_f \widehat{\mathcal{L}}(f_t)$
10: **end for**
11: **Output**: the classification model f_T

This loss is differentiable and avoids the surrogate approximations (e.g., Sigmoid function) employed in prior work for conformal training. As we show in the next subsection, this rank-weighted loss not only aligns closely with the conformal objective but also admits provable guarantees.

We optimize the objective in Eq. (5) using stochastic gradient descent. The optimization algorithm is summarized in Algorithm 1. At each iteration t, we first sample a mini-batch $\mathcal{B}_t$ and compute the softmax scores $f_t(X_j)$ for each sample in this batch, using the current model f_t (see Line 4 and 5). Then, we compute the true label rank $R_{f_t}(X_j, Y_j)$ and cross-entropy loss $\ell_{f_t}(X_j, Y_j)$, respectively (see Line 6 and 7). Next, we compute the RWCE loss $\widehat{\mathcal{L}}(f_t)$ (see Line 8). Finally, we update the model parameters f_{t+1} (see Line 9). Once training is finished, we output the trained classification model f_T (Line 11). This procedure is simple, efficient (with only $\mathcal{O}(K)$ cost to compute ranks), and introduces no additional hyperparameters such as temperature or smoothing terms in surrogate-based objectives.

4.2 Learning Bounds of Proposed Objective

We now provide theoretical justifications for using the RWCE loss. Our analysis proceeds in three parts: (i) we show that the expected rank upper bounds the expected size of conformal prediction sets, (ii) we show that our proposed objective upper bounds the expected rank under mild conditions, and (iii) we show that our empirical loss generalizes well with standard concentration bounds.

We begin by formalizing the connection between the true-label rank and the prediction set size. The following theorem shows that, under any conformity score and calibration procedure, the expected size of the conformal prediction set is upper bounded by the expected rank of the true label, up to a small calibration slack:

Theorem 1. *For any nonconformity score S and coverage level $(1-\alpha) \in (0, 1)$, the following holds:*

$$\mathbb{E}[|\widehat{C}(X)|] \leq \mathbb{E}[R(X, Y)] + K\left(1 - \alpha + \frac{1}{n+1}\right). \tag{6}$$

Remark 1. *Theorem 1 implies that reducing the average rank directly tightens the prediction set, with an $\mathcal{O}(K/n)$ calibration slack, which is a model-independent constant. Although this slack may be numerically large, it does not depend on the classifier and cannot be reduced through training. Thus, the only optimizable part in the bound, expected rank, is a* sound surrogate *for expected set size.*

However, directly minimizing the expected rank is still not ideal, as it is a discrete, non-differentiable quantity. Instead, we minimize our proposed objective: a rank-weighted version of the cross-entropy loss. To show this objective is sound, we now prove that, under a mild condition, it upper bounds the expected rank (up to a constant shift that does not affect optimization). This bridges the gap between the desired rank minimization and the tractable training loss.

Theorem 2. *Suppose that $\mathbb{E}[(R(X,Y)-1)(\ell(X,Y)-1)] \geq -\mathbb{E}[\ell(X,Y)]$ or ℓ is the cross-entropy loss. Then the following inequality holds:*

$$\mathbb{E}[R(X,Y)-1] \leq \mathbb{E}[R(X,Y)\cdot\ell(X,Y)] = \mathcal{L}(f).$$

Remark 2. *This result shows that our objective $\mathcal{L}(f)$ serves as an upper bound on the expected rank (up to a constant* 1*), meaning that optimizing the RWCE loss leads to models with low expected rank and therefore small prediction sets by Theorem 1. Importantly, the additive constant* 1 *does not influence learning because it is independent of model parameters. Thus, our method can be viewed as optimizing a differentiable, rank-weighted variant of the cross-entropy loss. The covariance condition is mild and usually holds in practice. It says that the rank prediction is positively correlated to the CE loss. We validate this empirically in Figure 3.*

Theorem 1 establishes a fundamental connection between the expected prediction set size and the expected rank of the true label. This result formalizes the intuition that the more confidently a model scores the correct class (i.e., assigns it a high rank), the fewer classes are needed to meet the conformal coverage guarantee. In practice, this means that by designing models that minimize expected rank, we are implicitly driving down the prediction set size to achieve better predictive efficiency.

Theorem 2 strengthens this insight by showing that RWCE loss function directly upper bounds the expected rank under a mild and realistic correlation condition. This result provides strong theoretical justification for our proposed training objective: not only does it approximate the desired metric (expected rank), it does so in a provably tight and optimizable way.

Finally, to ensure that our proposed training objective is learnable from finite data, we establish a standard generalization bound. This shows that the empirical version of our loss concentrates tightly around its population counterpart:

Corollary 1. *(Generalization error bound, (Mohri, Rostamizadeh, and Talwalkar 2018))*

$$sup_{f\in\mathcal{H}}\left(\mathcal{L}(f)-\widehat{\mathcal{L}}(f)\right) \leq \sqrt{\frac{\log(1/\delta)}{n}}.$$

Remark 3. *This result ensures that minimizing the empirical loss $\widehat{\mathcal{L}}(f)$ effectively minimizes the population loss $\mathcal{L}(f)$ with high probability, even in finite-sample settings. Together with Theorems 1 and 2, this completes our theoretical justification for using RWCE for conformal training.*

5 Experiments

5.1 Experiment Setup

Datasets. We firstly evaluate our method on three widely used image classification benchmarks with varying levels of label granularity and class imbalance: CIFAR-100 (Krizhevsky, Hinton et al. 2009), Caltech-101 (Fei-Fei, Fergus, and Perona 2006), and iNaturalist (Van Horn et al. 2018). Each dataset is randomly partitioned into training, calibration, and testing subsets following standard conformal prediction procedures. Table 2 in Appendix summarizes the key statistics, including the number of classes, sample counts, and imbalance ratios. We also conduct experiments on SST-5 (Socher et al. 2013), a text classification benchmarks.

Backbone Models. We adopt two widely used convolutional architectures, ResNet (He et al. 2016) and DenseNet (Huang et al. 2017), as deep models in our experiments for image classification. Detailed hyperparameters for each configuration are provided in Appendix, Table 3. For SST-5, we employed the BERT (Devlin et al. 2019) model.

Nonconformity Scores. We adopt four widely used nonconformity scoring functions in conformal prediction: HPS (Vovk, Gammerman, and Shafer 2005; Lei, Robins, and Wasserman 2013), APS (Romano, Sesia, and Candes 2020), RAPS (Angelopoulos et al. 2020b), and SAPS (Huang et al. 2023). After training, each model is evaluated using HPS, APS, RAPS and SAPS based on the average prediction set size (APSS) at a fixed marginal coverage of 90%. This multi-score evaluation framework provides a robust assessment of how well the model generalizes under different conformity measures. A detailed overview of these scoring functions is provided in Appendix C.

Baseline Methods. We compare our approach against three representative baseline training methods: (i) CE, which performs standard empirical risk minimization using the cross-entropy loss without any uncertainty-aware regularization; (ii) CUT (Einbinder et al. 2022), which penalizes the discrepancy between the empirical cumulative distribution function (CDF) of nonconformity scores and the uniform distribution to mitigate model overconfidence; and (iii) ConfTr (Stutz et al. 2021), which directly trains models to produce calibrated prediction sets by minimizing a differentiable approximation of the average set size. Specifically, it employs a Sigmoid approximation of the indicator function to define a differentiable loss that jointly penalizes large prediction sets and miscoverage. For CUT and ConfTr, we use the HPS score for training. The details of CUT and ConfTr are in Appendix C.

Evaluation Metrics. Our first evaluation metric is the average prediction set size (APSS), defined as $\text{APSS} = \frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{i\in\mathcal{D}_{\text{test}}}|C_f(X_i)|$. In addition to efficiency, coverage is a fundamental metric in conformal prediction. Accordingly, we report marginal coverage (Marg-Cov) as

Model	HPS				APS			
	CE	CUT	ConfTr	RWCE	CE	CUT	ConfTr	RWCE
Caltech-101								
ResNet	1.52 ± 0.045	1.55 ± 0.039	1.26 ± 0.03	**0.96** $\pm$ **0.008** ($\downarrow$ 23.81%)	4.96 ± 0.094	4.89 ± 0.095	4.25 ± 0.081	**1.33** $\pm$ **0.017** ($\downarrow$ 68.71%)
DenseNet	3.51 ± 0.10	1.66 ± 0.038	3.13 ± 0.07	**0.94** $\pm$ **0.005** ($\downarrow$ 43.37%)	8.92 ± 0.18	4.60 ± 0.078	9.69 ± 0.22	**1.27** $\pm$ **0.008** ($\downarrow$ 72.39%)
CIFAR-100								
ResNet	3.39 ± 0.10	2.91 ± 0.08	3.98 ± 0.077	**2.68** $\pm$ **0.083** ($\downarrow$ 7.90%)	3.98 ± 0.13	3.49 ± 0.104	5.13 ± 0.117	**3.08** $\pm$ **0.078** ($\downarrow$ 11.75%)
DenseNet	2.59 ± 0.053	2.07 ± 0.06	2.19 ± 0.034	**2.04** $\pm$**0.051** ($\downarrow$ 1.45%)	3.38 ± 0.12	**2.19** $\pm$ **0.060**	3.04 ± 0.069	2.335 ± 0.076 ($\uparrow$ 6.62%)
iNaturalist								
ResNet	98.69 ± 8.86	73.19 ± 3.13	76.30 ± 3.12	**69.05** $\pm$ **2.56** ($\downarrow$ 5.66%)	95.18 ± 3.50	79.58 ± 2.87	87.80 ± 1.97	**73.39** $\pm$**2.32** ($\downarrow$ 7.78%)
DenseNet	93.18 ± 2.92	74.11 ± 2.35	72.25 ± 2.07	**67.17** $\pm$ **2.07** ($\downarrow$ 7.03%)	101.55 ± 2.97	87.27 ± 2.30	92.88 ± 2.89	**75.65** $\pm$ **2.44** ($\downarrow$ 13.31%)

Table 1: Overall comparison on three datasets and calibrated by HPS and APS on DenseNet and ResNet with $\alpha = 0.1$. All models are evaluated under both HPS and APS calibration strategies to assess robustness across scoring functions. We report the mean and standard deviation of the reported APSS values over 10 independent runs. We benchmark four methods: standard CE, CUT, ConfTr, and RWCE. Arrows $\downarrow$ and $\uparrow$ indicate improvements or degradations in predictive efficiency relative to the best baseline. Overall, RWCE consistently produces the smallest prediction sets across all datasets and evaluation metrics, demonstrating a relative improvement of 21.38% in average prediction set size—highlighting its superior calibration quality and generalization capability.

our second evaluation criterion, defined as Marg-Cov $= \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{i \in \mathcal{D}_{\text{test}}} \mathbb{1}[Y_i \in C_f(X_i)]$.

5.2 Results and Discussion

RWCE generates smaller prediction sets. Table 1 reports the APSS of all methods across three datasets under a fixed coverage level of $\alpha = 0.1$. All models are evaluated using both HPS and APS scoring functions during the evaluation phase, with details provided in Appendix D. Our method, RWCE, demonstrates consistently strong performance across all datasets and evaluation settings by producing smaller and more efficient prediction sets than existing baselines. Averaging the relative improvements across all configurations, RWCE achieves a 21.38% reduction in prediction set size compared to the strongest baseline in each case. Under HPS evaluation, RWCE consistently achieves the smallest APSS across all dataset–model combinations. The gains are even more pronounced under APS evaluation, further reinforcing RWCE's effectiveness in producing compact and well-calibrated prediction sets—except for a slight degradation on CIFAR-100 with DenseNet, where RWCE still remains competitive. For completeness, coverage statistics and additional experiments using RAPS and SAPS scoring functions are reported in Appendix E. Additional experiments on SST-5 with the BERT model are also reported in Appendix E. We further visualize the APSS comparison of the three methods on CIFAR-100 in Figure 2 (c), using both ResNet and DenseNet evaluated by the HPS score. The figure clearly demonstrates that RWCE yields significantly smaller prediction sets after each method converges.

RWCE Converges Stably During Training. To assess the optimization dynamics of RWCE, we visualize its training loss (according to remark 2) over 60 fine-tuning epochs on CIFAR-100 using both ResNet and DenseNet architectures, as shown in Figure 2(a). The training curves exhibit a consistent convergence pattern across both models. Specifically, RWCE experiences some oscillations in the early epochs—particularly visible in ResNet where multiple sharp peaks occur within the first 25 epochs—followed by a sharp decline between epoch 25 and 30. After this point, the loss stabilizes rapidly and converges smoothly by around epoch 40. This loss corresponds to a rank-weighted objective based on the rank of the true label, which we theoretically show to be a tight upper bound on the expected prediction set size. The observed convergence behavior thus not only indicates optimization stability, but also reflects the ability of RWCE to directly approximate the set size minimization objective without relying on relaxed approximations of the indicator function.

RWCE Directly Minimizes Prediction Set Size. To examine whether RWCE effectively minimizes the true prediction set size, we compare its training loss with the actual APSS on the training set, as shown in Figure 2(b). For both ResNet and DenseNet, the training loss inflated with 1 (according to remark 2) closely upper bounds the APSS throughout training, with the two curves converging to similar values after epoch 40. This tight alignment empirically validates our theoretical claim that the RWCE objective serves as a provable upper bound on the expected set size. Furthermore, we track the test-time APSS of RWCE, ConfTr, and CUT under HPS calibration during training in Figure 2(c). While RWCE exhibits some fluctuation in the early stages of training—especially on the ResNet backbone—it rapidly stabilizes and consistently achieves the smallest prediction sets in the later epochs across both model architectures. This highlights its superior efficiency in directly minimizing prediction set size. Although ConfTr converges more smoothly, its final prediction sets remain larger than those of RWCE, which may reflect the limitations of optimizing a soft relaxation of the true objective. Together, these results suggest that RWCE not only aligns well with the evaluation metric but also enables the construction of compact and calibrated prediction sets without relying on soft indicators.

Assumption in Theorem 2 is empirically valid. To verify the practical validity of the alignment assumption required by Theorem 2, we compute and track both sides

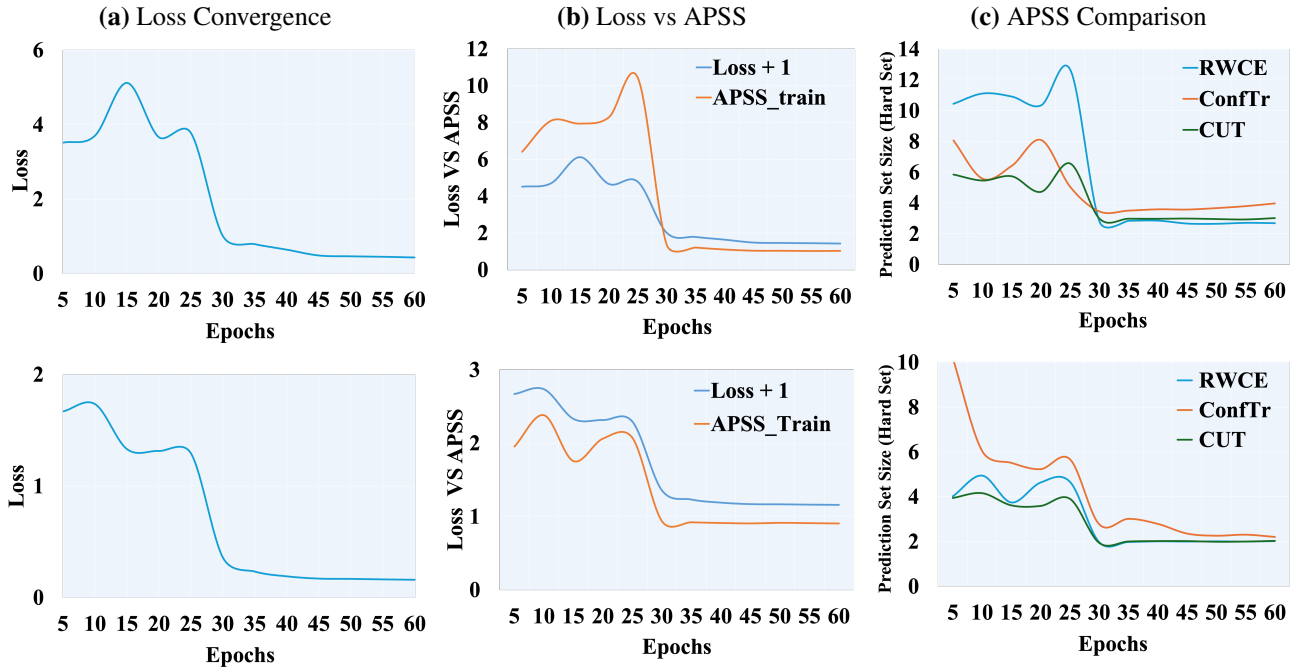

Figure 2: Justification experiments using ResNet (Top row) and DenseNet (Bottom row) on CIFAR-100. **(a)** the training loss convergence of RWCE. The results demonstrate that RWCE converges smoothly and stably on both architectures. **(b)** the RWCE training loss inflated by 1 (according to remark 2) with the actual APSS on the validation set. The loss closely upper bounds APSS with a small and stable gap, indicating that RWCE effectively approximates and directly minimizes the true set size objective. **(c)** the APSS of RWCE, ConfTr, and CUT calibrated by HPS score. RWCE consistently achieves smaller prediction sets on both architectures, confirming its superior efficiency in directly minimizing set size.

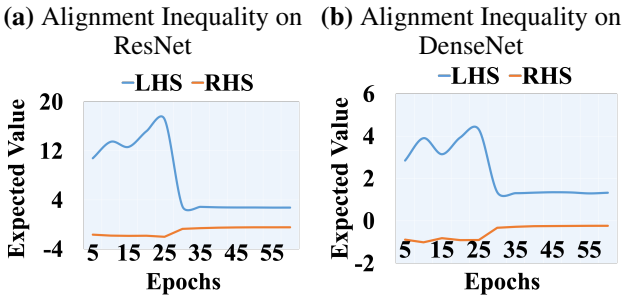

Figure 3: Empirical validation of LHS ($\mathbb{E}[(R(X,Y)-1)(\ell(X,Y)-1)]$) and RHS ($-\mathbb{E}[\ell(X,Y)]$) for the alignment assumption in Theorem 2, which states that $\mathbb{E}[(R(X,Y)-1)(\ell(X,Y)-1)] \geq -\mathbb{E}[\ell(X,Y)]$ during training on CIFAR-100 using ResNet (a) and DenseNet (b). We plot the left-hand side (expected rank–loss interaction term, blue) and right-hand side (negative expected cross-entropy loss, orange). In both cases, the blue curve remains significantly above the orange curve throughout training, confirming that the inequality holds in practice.

of the inequality throughout training. Specifically, we compare the left-hand side $\mathbb{E}[(R(X,Y)-1)(\ell(X,Y)-1)]$ and the right-hand side $-\mathbb{E}[\ell(X,Y)]$. Figure 3 shows the two curves on the CIFAR-100 training set using both ResNet and DenseNet architectures across all epochs. In both models, we observe that the expected rank–loss interaction term (blue curve, i.e., $\mathbb{E}[(R(X,Y)-1)(\ell(X,Y)-1)]$) consistently exceeds the magnitude of the average negative cross-entropy loss (orange curve, i.e., $-\mathbb{E}[\ell(X,Y)]$) throughout training. This confirms that the inequality assumed in Theorem 2 holds empirically. These results provide strong support for the theoretical justification of our objective, demonstrating that the rank-weighted loss meaningfully upper bounds the expected rank in real-world training scenarios.

6 Conclusion

In this paper, we proposed RWCE, a simple and theoretically grounded conformal training method that avoids the use of surrogate indicator approximations. Instead, RWCE minimizes a rank-weighted cross-entropy loss, where each training sample is reweighted according to the rank of its true label. We provide rigorous theoretical analysis showing that this objective tightly upper bounds the expected prediction set size and admits favorable generalization guarantees under mild conditions. Empirical results on benchmark datasets confirm the effectiveness of RWCE: it consistently reduces prediction set size compared to prior conformal training methods with an average 21.38% reduction, while preserving valid coverage. Our approach demonstrates a principled and practical alternative for conformal prediction, with implications for building adaptive and efficient uncertainty quantification methods.

Acknowledgements

The authors gratefully acknowledge the in part support by the USDA-NIFA funded AgAID Institute award 2021-67021-35344, and the NSF grant CNS-2312125, IIS-2443828, DUE-2519063. The views expressed are those of the authors and do not reflect the official policy or position of the USDA-NIFA and NSF.

References

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76: 243–297.

Angelopoulos, A.; Bates, S.; Malik, J.; and Jordan, M. I. 2020a. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*.

Angelopoulos, A. N.; and Bates, S. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Angelopoulos, A. N.; Bates, S.; Jordan, M.; and Malik, J. 2021. Uncertainty Sets for Image Classifiers using Conformal Prediction. In *International Conference on Learning Representations*.

Angelopoulos, A. N.; Bates, S.; Malik, J.; and Jordan, M. I. 2020b. Uncertainty Sets for Image Classifiers Using Conformal Prediction. *arXiv preprint arXiv:2009.14193*.

Babbar, V.; Bhatt, U.; and Weller, A. 2022. On the utility of prediction sets in human-ai teams. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*.

Bachute, M. R.; and Subhedar, J. M. 2021. Autonomous driving architectures: insights of machine learning and deep learning algorithms. *Machine Learning with Applications*, 6: 100164.

Bao, J.; Dang, C.; Luo, R.; Zhang, H.; and Zhou, Z. 2025. Enhancing Adversarial Robustness with Conformal Prediction: A Framework for Guaranteed Model Reliability. In *Forty-second International Conference on Machine Learning*. PMLR.

Bates, S.; Angelopoulos, A.; Lei, L.; Malik, J.; and Jordan, M. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*, 68(6): 1–34.

Begoli, E.; Bhattacharya, T.; and Kusnezov, D. 2019. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1): 20–23.

Bellotti, A. 2021. Optimized conformal classification using gradient descent approximation. *arXiv preprint arXiv:2105.11255*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

Ding, T.; Angelopoulos, A.; Bates, S.; Jordan, M.; and Tibshirani, R. J. 2024. Class-conditional conformal prediction with many classes. *Advances in Neural Information Processing Systems*, 36.

Einbinder, B.-S.; Romano, Y.; Sesia, M.; and Zhou, Y. 2022. Training uncertainty-aware classifiers with conformalized deep learning. *Advances in Neural Information Processing Systems*, 35: 22380–22395.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, 28(4): 594–611.

Fisch, A.; Schuster, T.; Jaakkola, T.; and Barzilay, R. 2021a. Efficient Conformal Prediction via Cascaded Inference with Expanded Admission. In *Proceedings of The Tenth International Conference on Learning Representations*.

Fisch, A.; Schuster, T.; Jaakkola, T.; and Barzilay, R. 2021b. Few-shot conformal prediction with auxiliary tasks. In *International Conference on Machine Learning*, 3329–3339. PMLR.

Ghosh, S.; Shi, Y.; Belkhouja, T.; Yan, Y.; Doppa, J.; and Jones, B. 2023. Probabilistically robust conformal prediction. In *Uncertainty in Artificial Intelligence*, 681–690. PMLR.

Gibbs, I.; and Candes, E. 2021. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34: 1660–1672.

Gibbs, I.; Cherian, J. J.; and Candès, E. J. 2023. Conformal Prediction With Conditional Guarantees. *arXiv preprint arXiv:2305.12616*.

Guan, L. 2023. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1): 33–50.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, J.; Xi, H.; Zhang, L.; Yao, H.; Qiu, Y.; and Wei, H. 2024. Conformal Prediction for Deep Classifier via Label Ranking. In *International Conference on Machine Learning*, 20331–20347. PMLR.

Huang, K.; Jin, Y.; Candès, E.; and Leskovec, J. 2023. Uncertainty Quantification over Graph with Conformalized Graph Neural Networks. In *NeurIPS 2023*.

Humbert, P.; Le Bars, B.; Bellet, A.; and Arlot, S. 2023. One-shot federated conformal prediction. In *International Conference on Machine Learning*, 14153–14177. PMLR.

Kiyani, S.; Pappas, G. J.; and Hassani, H. 2024. Length Optimization in Conformal Prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lei, J.; Robins, J.; and Wasserman, L. 2013. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501): 278–287.

Lu, C.; Angelopoulos, A. N.; and Pomerantz, S. 2022. Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 545–554. Springer.

Lu, C.; Lemay, A.; Chang, K.; Höbel, K.; and Kalpathy-Cramer, J. 2022. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 12008–12016.

Lu, C.; Yu, Y.; Karimireddy, S. P.; Jordan, M.; and Raskar, R. 2023. Federated conformal predictors for distributed uncertainty quantification. In *International Conference on Machine Learning*, 22942–22964. PMLR.

Luo, R.; and Zhou, Z. 2025. Reliable classification through rank-based conformal prediction sets. *Pattern Recognition*, 112330.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2018. *Foundations of machine learning*. MIT press.

Plassier, V.; Kotelevskii, N.; Rubashevskii, A.; Noskov, F.; Velikanov, M.; Fishkov, A.; Horvath, S.; Takac, M.; Moulines, E.; and Panov, M. 2024. Efficient conformal prediction under data heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, 4879–4887. PMLR.

Plassier, V.; Makni, M.; Rubashevskii, A.; Moulines, E.; and Panov, M. 2023. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, 27907–27947. PMLR.

Romano, Y.; Patterson, E.; and Candes, E. 2019. Conformalized quantile regression. *Advances in neural information processing systems*, 32.

Romano, Y.; Sesia, M.; and Candes, E. 2020. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33: 3581–3591.

Sadinle, M.; Lei, J.; and Wasserman, L. 2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234.

Shafaei, S.; Kugele, S.; Osman, M. H.; and Knoll, A. 2018. Uncertainty in machine learning: A safety perspective on autonomous driving. In *Computer Safety, Reliability, and Security: SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings 37*, 458–464. Springer.

Shafer, G.; and Vovk, V. 2008. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, 9(3).

Shi, Y.; GHOSH, S.; Belkhouja, T.; Doppa, J.; and Yan, Y. 2024. Conformal Prediction for Class-wise Coverage via Augmented Label Rank Calibration.

Shi, Y.; Shahrokhi, H.; Jia, X.; Chen, X.; Doppa, J. R.; and Yan, Y. 2025. Direct Prediction Set Minimization via Bilevel Conformal Classifier Training. *arXiv preprint arXiv:2506.06599*.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

Straitouri, E.; Wang, L.; Okati, N.; and Rodriguez, M. G. 2023. Improving Expert Predictions with Conformal Prediction. In *International Conference on Machine Learning (ICML)*.

Stutz, D.; Cemgil, A. T.; Doucet, A.; et al. 2021. Learning optimal conformal classifiers. *arXiv preprint arXiv:2110.09192*.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8769–8778.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*. Springer Science & Business Media.

Xi, H.; Liu, K.; Zeng, H.; Sun, W.; and Wei, H. 2025. Robust Online Conformal Prediction under Uniform Label Noise. *CoRR*.

Yan, G.; Romano, Y.; and Weng, T.-W. 2024. Provably Robust Conformal Prediction with Improved Efficiency. arXiv:2404.19651.

Yang, S.; and Fevens, T. 2021. Uncertainty quantification and estimation in medical image classification. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part III 30*, 671–683. Springer.

Zhu, Y.; Hernández, D.; He, Y.; Ding, Z.; Xiong, B.; Kharlamov, E.; and Staab, S. 2025. Predicate-Conditional Conformalized Answer Sets for Knowledge Graph Embeddings. *arXiv preprint arXiv:2505.16877*.