Conformal Prediction for Deep Classifier via Label Ranking

Jianguo Huang [*1 2 †] **Huajun Xi** [*1] **Linjun Zhang** [3] **Huaxiu Yao** [4] **Yue Qiu** [5] **Hongxin Wei** [1]

Abstract

Conformal prediction is a statistical framework that generates prediction sets containing ground-truth labels with a desired coverage guarantee. The predicted probabilities produced by machine learning models are generally miscalibrated, leading to large prediction sets in conformal prediction. To address this issue, we propose a novel algorithm named *Sorted Adaptive Prediction Sets* (SAPS), which discards all the probability values except for the maximum softmax probability. The key idea behind SAPS is to minimize the dependence of the non-conformity score on the probability values while retaining the uncertainty information. In this manner, SAPS can produce compact prediction sets and communicate instance-wise uncertainty. Extensive experiments validate that SAPS not only lessens the prediction sets but also broadly enhances the conditional coverage rate of prediction sets.

1. Introduction

Machine learning is being deployed in many high-stakes tasks, such as autonomous driving (Bojarski et al., 2016), medical diagnostics (Caruana et al., 2015), and financial decision-making. The trust and safety in these applications are critical, as any erroneous prediction can be costly and dangerous. To assess the reliability of predictions, a popular solution is to quantify the model uncertainty, such as confidence calibration (Guo et al., 2017), MC-Dropout (Gal & Ghahramani, 2016), and Bayesian neural network (Smith, 2013). However, these methods lack theoretical guarantees of the model performance. This gives rise to the importance of Conformal Prediction (CP) (Vovk et al., 2005; Shafer & Vovk, 2008; Balasubramanian et al., 2014; Angelopoulos & Bates, 2021), which yields prediction sets containing ground-truth labels with a desired coverage guarantee.

To determine the size of the prediction set, CP algorithms generally design non-conformity scores to quantify the deviation degree between a new instance and the training data. A higher non-conformity score is associated with a larger prediction set or region, indicating a lower confidence level in the prediction. For example, Adaptive Prediction Sets (APS) (Romano et al., 2020) calculates the score by accumulating the sorted softmax values in descending order. However, the softmax probabilities typically exhibit a long-tailed distribution, allowing for easy inclusion of those tail classes in the prediction sets. To alleviate this issue, Regularized Adaptive Prediction Sets (RAPS) (Angelopoulos et al., 2021) excludes unlikely classes by appending a penalty to classes beyond a specified threshold. Yet, the non-conformity score of RAPS involves unreliable softmax probabilities, leading to suboptimal performance in conformal prediction. This motivates our question: *does the probability value play a critical role in conformal prediction?*

In this work, we show that the value of softmax probability might be redundant information for conformal prediction. We provide an empirical analysis by removing the exact value of softmax probability while preserving the relative rankings of labels. The results indicate that APS without probability value yields much more compact prediction sets than APS using the softmax outputs, at the same coverage rate. Theoretically, we show that, by removing the probability value, the size of prediction sets generated by APS is consistent with the model prediction accuracy. In other words, a model with higher accuracy can produce smaller prediction sets, using APS without access to the probability value. The details of the analysis are presented in Section 3.1.

Inspired by the analysis, our key idea is to minimize the dependence of the non-conformity score on the probability values, while retaining the uncertainty information. Specifically, we propose the *Sorted Adaptive prediction sets* (dubbed **SAPS**), which discards all the probability values except for the maximum softmax probability in the construc-

[*]Equal contribution ,[†]Work done while working at SUSTech as a visiting scholar. [1]Department of Statistics and Data Science, Southern University of Science and Technology. [2]The School of Information Science and Technology, ShanghaiTech University [3]Department of Statistics, Rutgers University [4]Department of Computer Science, University of North Carolina at Chapel Hill [5]College of Mathematics and Statistics, Chongqing University. Correspondence to: Hongxin Wei <weihx@sustech.edu.cn>, Yue Qiu <qiuyue@cqu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

tion of non-conformity score. After sorting in descending order, this can be achieved by replacing the non-maximum probability values with a constant. In effect, SAPS can not only produce sets of small size but also communicate instance-wise uncertainty. In addition, we show that as the constant approaches infinity, the expected value of set size from SAPS is asymptotically equivalent to that of APS without probability value (see Figure 3b).

To verify the effectiveness of our method, we conduct thorough empirical evaluations on common benchmarks, including CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and ImageNet (Deng et al., 2009). The results demonstrate that SAPS is superior to the compared methods, including APS and RAPS. For example, our approach reduces the average size of prediction sets from 20.95 to 2.98 – only $\frac{1}{7}$ of the prediction set size from APS. Compared to RAPS, we show that SAPS produces a higher conditional coverage rate and exhibits better adaptability to the instance difficulty. Code are publicly available at https://github.com/ml-stat-Sustech/conformal_prediction_via_label_ranking.

We summarize our contributions as follows:

1. We show that the probability value might be unnecessary for APS. Specifically, APS without probability value generates smaller prediction sets than vanilla APS. Theoretically, we show that this variant of APS can provide stable prediction sets, in which the set size is consistent with the prediction accuracy of models.

2. We propose a novel non-conformity score – SAPS that minimizes the dependency on probability value while retaining the uncertainty information.

3. Extensive experimental results demonstrate the effectiveness of our proposed method. We show that SAPS lessens the prediction sets and broadly enhances the conditional coverage rate.

4. We provide analyses to improve our understanding of the proposed method. In particular, we validate that both retaining the maximum softmax probability and confidence calibration are important for SAPS to generate compact prediction sets (see Section 5).

2. Preliminaries

In this work, we consider the multi-class classification task with K classes. Let $\mathcal{X} \subset \mathbb{R}^d$ be the input space and $\mathcal{Y} := \{1, \ldots, K\}$ be the label space. We use $\hat{\pi} : \mathcal{X} \to \mathbb{R}^K$ to denote the pre-trained neural network that is used to predict the label of a test instance. Let $(X, Y) \sim \mathcal{P}_{\mathcal{X}\mathcal{Y}}$ denote a random data pair satisfying a joint distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$. Ideally, $\hat{\pi}_y(\boldsymbol{x})$ can be used to approximate the conditional probability of the class y given an image feature $\boldsymbol{x}$, i.e., $\mathbb{P}[Y = y|X = \boldsymbol{x}]$. Then, the model prediction in classification tasks is generally made as: $\hat{y} = \arg\max_{y \in \mathcal{Y}} \hat{\pi}_y(\boldsymbol{x})$.

Conformal prediction. To provide a formal guarantee for the model performance, conformal prediction (Vovk et al., 2005) is designed to produce prediction sets containing ground-truth labels with a desired probability. Instead of predicting one-hot labels from the model outputs, the goal of conformal prediction is to construct a set-valued mapping $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$, which satisfies the *marginal coverage*:

$$\mathbb{P}(Y \in \mathcal{C}(X)) \geq 1 - \alpha, \tag{1}$$

where $\alpha \in (0, 1)$ denotes the desired error rate and $\mathcal{C}(X)$ is a subset of $\mathcal{Y}$.

Before deployment, conformal prediction begins with a calibration step, using the calibration set $\mathcal{D}_{cal} := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. The samples of the calibration set are i.i.d., having been drawn from the distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$. Specifically, we calculate a non-conformity score $s_i = S(\boldsymbol{x}_i, y_i)$ for each example $(\boldsymbol{x}_i, y_i)$ in the calibration set, where s_i measures the degree of deviation between the given example and the training data. The $1 - \alpha$ quantile of the non-conformity scores $\{s_i\}_{i=1}^n$ is then determined as a threshold τ. Formally, the value of τ can be obtained as shown below:

$$\tau = \inf\{s : \frac{|\{i \in \{1, \ldots, n\} : s_i \leq s\}|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\}. \tag{2}$$

During testing, we calculate the non-conformity score for each label given a new instance $\boldsymbol{x}_{n+1}$. Then, the corresponding prediction set $\mathcal{C}(\boldsymbol{x}_{n+1})$ comprises possible labels whose non-conformity score $S(\boldsymbol{x}_{n+1}, y)$ falls within $\hat{\tau}$:

$$\mathcal{C}_{1-\alpha}(\boldsymbol{x}_{n+1}; \tau) := \{y \in \mathcal{Y} : S(\boldsymbol{x}_{n+1}, y) \leq \tau\}. \tag{3}$$

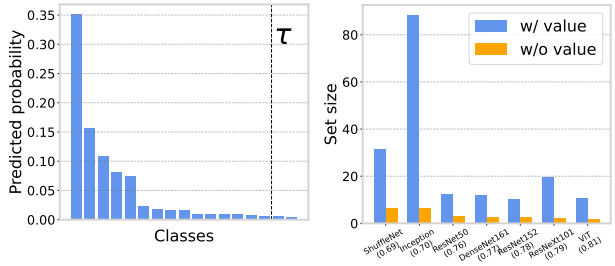
The above equation exhibits a nesting property of threshold, i.e.,

$$\tau_1 \leq \tau_2 \Longrightarrow \mathcal{C}_{1-\alpha}(\boldsymbol{x}_{n+1}; \tau_1) \subseteq \mathcal{C}_{1-\alpha}(\boldsymbol{x}_{n+1}; \tau_2). \tag{4}$$

This property shows that with a lower value of τ, the model tends to produce a smaller prediction set, indicating a higher level of confidence in the prediction. Conversely, the increase of τ will enlarge the size of the prediction set, suggesting greater uncertainty of the prediction. In this manner, conformal prediction can be used to estimate the uncertainty or reliability of the model's predictions.

Adaptive prediction sets (APS). In the APS method (Romano et al., 2020), the non-conformity scores are calculated by accumulating the softmax probabilities in descending order. Formally, given a data pair $(\boldsymbol{x}, y)$, the non-conformity score can be computed by:

$$S_{aps}(\boldsymbol{x}, y, u; \hat{\pi}) := \sum_{i=1}^{o(y, \hat{\pi}(\boldsymbol{x}))-1} \hat{\pi}_{(i)}(\boldsymbol{x}) + u \cdot \hat{\pi}_{(o(y, \hat{\pi}(\boldsymbol{x})))}(\boldsymbol{x}), \tag{5}$$



(a) (b)

Figure 1. (a) Softmax probabilities for an instance from ImageNet are arranged in descending order. (b) Set size on various models. We use "w/ value" and "w/o value" to represent the vanilla APS and APS without probability values, respectively. The numbers in brackets represent the accuracy of the model. The sizes of the prediction sets are decreased after removing the probability value.

where $o(y, \hat{\pi}(\boldsymbol{x}))$ denotes the index of $\hat{\pi}_y(\boldsymbol{x})$ in the sorted softmax probabilities, i.e., $\hat{\pi}_{(1)}(\boldsymbol{x}), \ldots, \hat{\pi}_{(K)}(\boldsymbol{x})$, and u is an independent random variable satisfying a uniform distribution on $[0, 1]$. Similarly, the prediction set of APS with the error rate α for a test instance $\boldsymbol{x}_{n+1}$ is given by

$$\mathcal{C}_{1-\alpha}(\boldsymbol{x}_{n+1}, u_{n+1}; \tau) := \{y \in \mathcal{Y} : S_{aps}(\boldsymbol{x}_{n+1}, y, u_{n+1}; \hat{\pi}) \leq \tau\}, \tag{6}$$

where $u_{n+1} \sim U[0, 1]$ and τ is defined by Equation 2. In particular, a larger value of τ will lead to a larger prediction set, representing higher uncertainty. Thus, the prediction set defined by Equation 6 satisfies the nesting property in Equation 4.

In addition, the prediction set $\mathcal{C}_{1-\alpha}(\boldsymbol{x}_{n+1}, u_{n+1}; \tau)$ with the τ obtained by Equation 2 has a finite-sample marginal coverage guarantee, as stated in the following theorem.

Theorem 2.1. *(Angelopoulos et al., 2021) Suppose the calibration data $(X_i, Y_i, U_i)_{i=1,\ldots,n}$ and a test instance $(X_{n+1}, Y_{n+1}, U_{n+1})$ are exchangeable. Let the set-valued function $\mathcal{C}_{1-\alpha}(\boldsymbol{x}, u; \tau)$ satisfy the nesting property of τ in Equation 4. For τ defined in Equation 2, we have the following coverage guarantee:*

$$P\left(Y_{n+1} \in \mathcal{C}_{1-\alpha}\left(X_{n+1}, U_{n+1}; \tau\right)\right) \geq 1 - \alpha.$$

In practice, the assumption of exchangeability in the above Theorem is promising as it is weaker than the i.i.d. assumption (Teng et al., 2023). However, the softmax probabilities $\hat{\pi}(\boldsymbol{x})$ typically exhibit a long-tailed distribution, where the tail probabilities with small values can be easily included in the prediction sets. Consequently, APS tends to produce large prediction sets for all inputs, regardless of the instance difficulty. For example, in Figure 1a, the long-tail probability distribution results in the non-conformity scores of many classes falling within the threshold τ. This motivates our analysis to investigate the role of probability values in conformal prediction.

3. Motivation and method

3.1. Motivation

To analyze the role of probability values, we perform an ablation study by removing the influence of probability values in Equation 5. In particular, after arranging softmax probabilities in descending order, we substitute them with a positive constant γ (e.g., $\gamma = 1$).

Formally, the modified non-conformity score for a data pair $(\boldsymbol{x}, y)$ with a pre-trained model $\hat{\pi}$ can be give by:

$$S_{cons}(\boldsymbol{x}, y, u; \hat{\pi}, \gamma) := \gamma \cdot [o(y, \hat{\pi}(\boldsymbol{x})) - 1 + u]. \tag{7}$$

In what follows, we thoroughly probe the impacts of γ. We start by giving the following definition.

Lemma 3.1. *(Vovk et al., 2005) Let $\mathcal{Z}$ be a space of examples. Given a significance level, two non-conformity score functions A and B could provide the same prediction sets if $A(z) \leq A(z') \Longleftrightarrow B(z) \leq B(z')$ for all $z, z' \in \mathcal{Z}$.*

Lemma 3.1 states that in conformal prediction, different non-conformity score functions are equivalent if they lead to the same non-conformity order for any data examples.

Proposition 3.2. *For all $\gamma_1, \gamma_2 > 0$, the score functions $S_{cons}(\boldsymbol{x}, y, u; \hat{\pi}, \gamma_1)$ and $S_{cons}(\boldsymbol{x}, y, u; \hat{\pi}, \gamma_2)$ could provide the same prediction sets in conformal prediction.*

The proof of the above proposition is presented in Appendix A. According to Proposition 3.2, we can find that the value of γ in $S_{cons}(\boldsymbol{x}, y, u; \hat{\pi}, \gamma)$ is irrelevant with the prediction sets. Thus, we fix the constant γ to 1 for simplification. We conduct experiments on ImageNet (Deng et al., 2009) to compare the new non-conformity score to the vanilla APS. Here, we set the desired error rate as 10%, i.e., $\alpha = 0.1$. Following previous works (Romano et al., 2019; Angelopoulos et al., 2021; Ghosh et al., 2023), we first randomly split the test dataset of ImageNet into two subsets: a conformal calibration subset of size 30K and a test subset of size 20K. For the network architecture, we use seven models trained on ImageNet, with different levels of prediction performance (see Figure 1b). All models are calibrated by the temperature scaling procedure (Guo et al., 2017) on the calibration dataset. Finally, experiments are repeated ten times and the median results are reported.

Probability values may not be necessary. Figure 1b presents the results on various models, using APS with/without the probability value. The results indicate that APS solely based on label ranking generates smaller prediction sets than the vanilla APS, across various models. For example, on the Inception model, removing the

probability values reduces the set size from 88.18 to 6.33. Similarly, when applying a transformer-based ViT model (Touvron et al., 2021), APS without probability values also obtains a smaller set size. This comparison suggests that the probability values might be redundant for non-conformity scores in conformal prediction. We further explore this by theoretically analyzing the benefits of discarding softmax probability values in APS.

A theoretical interpretation. The above empirical results demonstrate that the probability value is not a critical component of the non-conformity score for conformal prediction. Here, we provide a formal analysis of APS without probability value through the following theorem:

Theorem 3.3. *Let A_r denote the accuracy of the top r predictions for a trained model $\hat{\pi}$ on an infinite calibration set. Given a significance level α, there exists an integer k satisfying $A_k \geq 1-\alpha > A_{k-1}$. For any test instance $\boldsymbol{x} \sim \mathcal{P}_{\mathcal{X}}$ and an independent random variable $u \sim U[0,1]$, the size of the prediction set $\mathcal{C}_{1-\alpha}(\boldsymbol{x}, u)$ generated by APS without probability value can be obtained by*

$$|\mathcal{C}_{1-\alpha}(\boldsymbol{x}, u)| = \begin{cases} k, & \text{if } u < \dfrac{1-\alpha-A_{k-1}}{A_k - A_{k-1}}, \\ k-1, & \text{otherwise.} \end{cases} \tag{8}$$

The expected value of the set size can be given by

$$\mathbb{E}_{u\sim[0,1]}[|\mathcal{C}_{1-\alpha}(\boldsymbol{x}, u)|] = k - 1 + \frac{1-\alpha-A_{k-1}}{A_k - A_{k-1}}. \tag{9}$$

The proof of Theorem 3.3 can be found in Appendix B. This is consistent with the theoretical results of RAPS, which indicates that the size of prediction sets generated by RAPS with heavy regularization will be at most equal to k (Angelopoulos et al., 2021). As indicated by Equation 9, the prediction set size generated by APS without probability value is consistent with k. In other words, a higher model accuracy will lead to a smaller value of k, indicating smaller prediction sets. This argument is clearly supported by experimental results shown in Figure 1b. In particular, we observe that using APS without probability value, models with higher accuracy produce smaller prediction sets, while the vanilla APS does not exhibit this characteristic. For example, the vanilla APS yields larger prediction sets for ResNeXt101 than for ResNet152, despite ResNeXt101 achieving higher predictive accuracy than ResNet152. The analysis demonstrates the advantage of removing probability value in APS, via decreasing the sensitivity to tail probabilities.

3.2. Method

In the above analysis, we demonstrate that removing the probability values in APS can largely decrease the size of prediction sets. Yet, the expected value of the set size (shown in Equation 8) will oscillate between $k-1$ and k, using APS without the probability value. This implies a shortcoming of the modified non-conformity score in adaptation to instance-wise uncertainty, leading to redundant prediction sets for easy examples.

To alleviate this limitation, we propose a novel conformal prediction algorithm, named *Sorted Adaptive Prediction Sets* (SAPS). The key idea behind SAPS is to minimize the dependence of the non-conformity score on the probability values, while retaining the uncertainty information. In particular, we discard all the probability values except for the maximum softmax probability (MSP), which is usually used to measure the model confidence in the prediction (Hendrycks & Gimpel, 2016).

Formally, the non-conformity score of SAPS for a data pair $(\boldsymbol{x}, y)$ can be calculated as

$$S_{saps}(\boldsymbol{x}, y, u; \hat{\pi}) := \begin{cases} u \cdot \hat{\pi}_{max}(\boldsymbol{x}), & \text{if} \quad o(y, \hat{\pi}(\boldsymbol{x})) = 1, \\ \hat{\pi}_{max}(\boldsymbol{x}) + (o(y, \hat{\pi}(\boldsymbol{x})) - 2 + u) \cdot \lambda, & \text{else,} \end{cases} \tag{10}$$

where λ is a hyperparameter representing the weight of ranking information, $\hat{\pi}_{max}(\boldsymbol{x})$ denotes the maximum softmax probability and u is a uniform random variable.

During inference, we calculate the non-conformity score for each label. In particular, the score is equivalent to that of APS for the label with the maximum softmax probability, while the scores of the resting labels are composed of the MSP and the ranking information from 2 to $o(y, \hat{\pi}(\boldsymbol{x}))$, i.e., $(o(y, \hat{\pi}(\boldsymbol{x})) - 2 + u)$. In this way, the new score of SAPS retains instance-wise uncertainty and mitigates the undesirable influence of tail probabilities.

With the new non-conformity score, the prediction sets of SAPS can be established according to Equation 6, following APS and RAPS. Consequently, we can conclude that SAPS also satisfies the marginal coverage guarantee, as stated in Theorem 2.1. Moreover, We provide an ablation experiment to investigate the relation between the prediction sets and the ranking weight λ in Section 5.

4. Experiments

4.1. Experimental Setup

Datasets. We consider three prominent datasets in our experiments: ImageNet (Deng et al., 2009), CIFAR-100 and CIFAR-10 (Krizhevsky et al., 2009), which are common benchmarks for conformal prediction. In particular, for ImageNet, its test dataset of 50,000 images is divided, allocating 30,000 images to the calibration set and 20,000 images to the test set. For both CIFAR-100 and CIFAR-10, the associated test dataset of 10,000 images is uniformly divided into two subsets: a calibration set and a test set,

Table 1. Performance comparison of various methods with different error rates α. We conduct experiments on ImageNet and CIFAR-10/100 datasets. **Bold** numbers indicate optimal performance.

	$\alpha = 0.1$						$\alpha = 0.05$					
	Coverage			Size ↓			Coverage			Size ↓		
Datasets	APS	RAPS	SAPS	APS	RAPS	SAPS	APS	RAPS	SAPS	APS	RAPS	SAPS
ImageNet	0.899	0.900	0.900	20.95	3.29	**2.98**	0.949	0.950	0.950	44.67	8.57	**7.55**
CIFAR-100	0.899	0.900	0.899	7.88	2.99	**2.67**	0.950	0.949	0.949	13.74	6.42	**5.53**
CIFAR-10	0.899	0.900	0.898	1.97	1.79	**1.63**	0.950	0.950	0.950	2.54	2.39	**2.25**

each comprising 5,000 images.

Models. We employ twelve different classifiers, including nine standard classifiers, two transformer-based models, i.e., ViT (Dosovitskiy et al., 2020) and DeiT (Touvron et al., 2021), and a Vision-Language Model named CLIP (Radford et al., 2021). Aside from CLIP with zero-shot prediction capabilities, the remaining models are the models pre-trained on ImageNet. For CIFAR-10 and CIFAR-100, these models will be fine-tuned on the pre-trained models. Moreover, all classifiers are calibrated by the Temperature scaling procedure (Guo et al., 2017) before applying CP methods.

Conformal prediction algorithms. We compare the proposed method against APS (Romano et al., 2020) and RAPS (Angelopoulos et al., 2021). For methods that has hyper-parameters, we choose the hyper-parameter that achieves the smallest set size on a validation set, which is a subset of the calibration set. Specifically, we tune the regularization hyperparameter of RAPS in $\{0.001, 0.01, 0.1, 0.15, \ldots, 0.5\}$ and hyperparameter λ in $\{0.02, 0.05, 0.1, 0.15, \ldots, 0.6\}$ for SAPS. All experiments are conducted with ten trials, and the median results are reported.

Additionally, we provide the details about the calibration process of conformal prediction algorithms, as shown:

1. **Split:** we split the full calibration set into a validation set and a calibration set ($20 : 80$ in this work);
2. **Tuning:** we use the validation set to choose proper hyperparameters of non-conformity score function;
3. **Conformal Calibration:** we use the calibration set to calculate the threshold τ.

For CP methods with hyperparameters (i.e., RAPS and SAPS), we use the "Split" and "Tuning" steps to tune the hyperparameters. For APS, we use the full calibration set. Therefore, all methods have access to the same dataset for fair comparison.

Evaluation. The primary metrics used for the evaluation of prediction sets are set size (average length of prediction sets, small value means high efficiency) and marginal coverage rate (fraction of testing examples for which prediction sets contain the ground-truth labels). These two metrics can be formally represented as :

$$\text{Size} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} |\mathcal{C}(\boldsymbol{x}_i)|$$

$$\text{Coverage rate} = \frac{1}{N_{test}} \sum_{i=1}^{N_{test}} \mathbf{1}(y_i \in \mathcal{C}(\boldsymbol{x}_i))$$

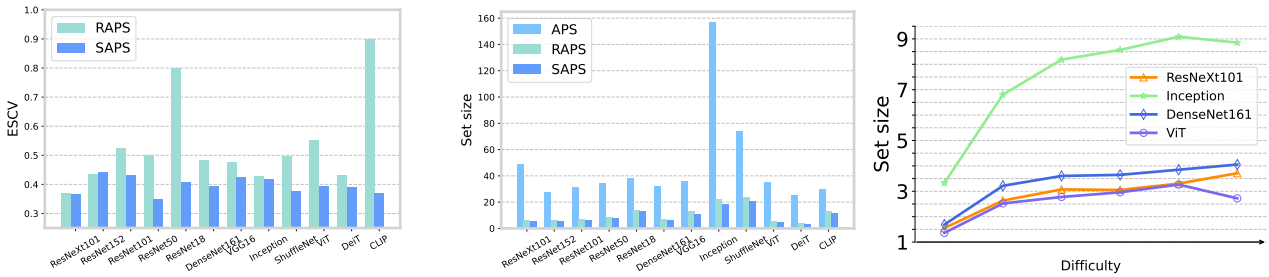
Conditional coverage rate. In this work, we propose an alternative metric to the *Size-stratified Coverage Violation* (Angelopoulos et al., 2021) named *Each-Size Coverage Violation* (ESCV), which is given by:

$$\text{ESCV}(\mathcal{C}, K) = \sup_j \max(0, 1 - \alpha - \tfrac{|\{i \in \mathcal{J}_j : y_i \in \mathcal{C}(\boldsymbol{x}_i)\}|}{|\mathcal{J}_j|})$$

where $\mathcal{J}_j = \{i : |\mathcal{C}(\boldsymbol{x}_i)| = j\}$ and $j \in \{1, \ldots, K\}$. Specifically, ESCV measures the most significant violation of prediction sets. This metric is practical because it only requires the set size, and is suitable for any classification problem, spanning from binary classes to large classes.

4.2. Results

SAPS generates smaller prediction sets. In Table 1, the performance of set sizes and coverage rates for various classification tasks are presented. We can observe that the coverage rate of all conformal prediction methods is close to the desired coverage $1 - \alpha$. At different significance levels (i.e., 0.1 and 0.05), the prediction set size is consistently reduced by SAPS for ImageNet, CIFAR-100 and CIFAR-10, compared to APS and RAPS. For example, when evaluated on ImageNet, SAPS reduces the average set size from 20.95 of APS to 2.98. Moreover, as the scale of the classification task increases, the efficiency improvement achieved by SAPS becomes increasingly evident. Overall, the experiments show that our method has the desired coverage rate and a smaller set size than APS and RAPS. Due to



(a) (b) (c)

Figure 2. (a) ESCV with different models on ImageNet with $\alpha = 0.1$. A good conformal prediction algorithm should keep the y-axis (e.g., ESCV) small. The results show that SAPS outperforms RAPS on most cases. (b) Set size on ImageNet-V2 at $\alpha = 0.1$. (c) Set size of various difficulties for multiple models on ImageNet. Small sets are required for easy examples, while hard ones require large sets.

space constraints, we only report the average results of multiple models on various classification tasks in Table 1, and detailed results for each model are available in Appendix C.

Experiments on distribution shifts. We also verify the effectiveness of our method on the new distribution, which is different from the training data distribution. Specifically, we divide the test dataset of ImageNet-V2 (Recht et al., 2019), which exhibits a distribution shift compared to the ImageNet, equally into a calibration set containing 5000 images and a test set containing 5000 images. Then, the test models are only pre-trained on ImageNet and not be fine-tuned. As shown in Figure 2b, the result shows that under $\alpha = 0.1$, our method can also generate the smallest sets when the conformal calibration set and the test set come from a new distribution.

SAPS acquires lower conditional coverage violation. In Figure 2a, we demonstrate that SAPS not only outperforms in efficiency but also boosts the conditional coverage rate, i.e., ESCV. Given that our study primarily focuses on improving the efficiency of prediction sets, the comparison of ESCV is exclusively conducted between SAPS and RAPS. The results, shown in Figure 2a, demonstrate that on ImageNet, SAPS would get smaller ESCV than RAPS for most models. For example, on CLIP, SAPS reduces the ESCV from 0.9 to 0.37. In addition, on ImageNet, we can observe that the ESCV of SAPS for different models is more stable than RAPS. Specifically, the ESCV of SAPS can keep a low value on most models, but in the case of RAPS, the maximum ESCV even gets 0.9. The detailed results on CIFAR-10 and CIFAR-100 are provided in Appendix D.

Moreover, to further confirm that SAPS achieves superior conditional coverage than RAPS, we report the results of *Size-stratified Coverage Violation* (SSCV). In this experiment, the coarse partitioning of the set sizes for SSCV is set as: 0-1, 2-3, 4-10, 11-100, and 101-1000, which follows the experimental setting of RAPS (Angelopoulos et al., 2021). The average SSCV of various models on three datasets are

Table 2. The average SSCV and ESCV of various models on three datasets. Bold numbers indicate optimal performance.

	ImageNet	ImageNet-V2	CIFAR100
Metrics	RAPS/SAPS(ours)		
SSCV↓	0.28/**0.23**	0.14/**0.13**	0.21/**0.18**
ESCV↓	0.53/**0.40**	0.43/**0.41**	0.37/**0.30**

shown in Table 2, with the experimental setup for ImageNet-V2 being consistent with the description provided in the preceding paragraph. We can observe that SAPS outperforms RAPS in conditional coverage. For instance, on ImageNet, SAPS records an SSCV of 0.23, which is 0.05 lower than that of RAPS. Although SAPS has better conditional coverage than RAPS, note that SSCV calculates the worst conditional coverage over the human-defined enclosures of set sizes, where different partitions lead to inconsistent SSCV results. Consequently, avoiding the human factor may lead to fair comparison, such as ESCV.

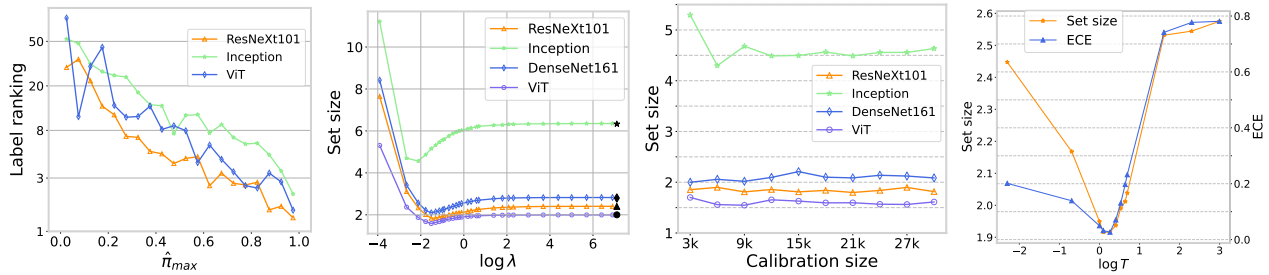
5. Discussion

SAPS exhibits adaptation. In the literature of conformal prediction, the size of prediction sets is expected to represent the inherent uncertainty of the classifier's predictions. Specifically, prediction sets should be larger for hard examples than for easy ones (Angelopoulos & Bates, 2021; Angelopoulos et al., 2021). In this work, we employ the rank of the ground-truth labels in the sorted softmax probabilities to denote the difficulty. For instance, examples with serious difficulty are assigned high ranks for their ground-truth labels. For simplicity, we choose a relatively coarse partitioning of image difficulty: 1, 2-3, 4-10, 11-100, and 101-1000. Then, we conduct an experiment on ImageNet to analyze the set size of different difficulties for SAPS.

In Figure 2c, a salient observation is that the set size rises

Table 3. Set size and ESCV for RAPS ($k_r = 1$) and SAPS. We report the average value across various models with $\alpha = 0.1$. The detailed results of each model are provided in the Appendix E. **Bold** numbers indicate optimal performance.

	Coverage		Size ↓		ESCV ↓	
Datasets	RAPS($k_r = 1$)	SAPS	RAPS($k_r = 1$)	SAPS	RAPS($k_r = 1$)	SAPS
ImageNet	0.900	0.900	3.24	**2.98**	0.631	**0.396**
CIFAR-100	0.899	0.899	2.79	**2.67**	0.390	**0.302**
CIFAR-10	0.900	0.898	**1.62**	1.63	0.138	**0.089**



(a) (b) (c) (d)

Figure 3. (a) The average ground-truth label ranking under different maximum softmax probabilities. Higher $\hat{\pi}_{max}$ have smaller label ranking. (b) Effect of the λ on set size across various models. The black markers (★, ♦, ▲, •) represent the results of APS without probability value. (c) Effect of the calibration dataset size on set size across various models. (d) Relationship between temperature and the set size of SAPS on ResNet152, where the horizon axis represents the log transformation of temperature T.

Table 4. Set size about retained top-k probability values on the ImageNet. We report the average set size across various models with $\alpha = 0.1$. **Bold** numbers indicate optimal performance.

	k				
Model	1	2	3	4	5
ResNeXt101	**1.80**	1.94	2.03	2.23	2.42
DenseNet161	**2.12**	2.16	2.27	2.46	2.65
ViT	**1.60**	1.67	1.76	1.86	1.97
Inception	4.60	**4.50**	4.55	4.77	4.95

with increasing difficulties for different models. For example, on the Inception model, SAPS enlarges the length of the prediction sets from approximately 3.5 for the simplest examples to nearly 9 for the hardest examples. Moreover, we notice that the classifier with a higher accuracy rate can have a smaller set size under different levels of difficulty compared to the classifier with a lower accuracy rate. For example, the ResNeXt101 model has a smaller set size than DenseNet161 across different difficulties.

Retained probability values. To mitigate the redundancy of prediction sets for easy examples, we retain MSP in the softmax probabilities, as shown in Equation 10. In this manner, the SAPS method can not only produce efficient prediction sets but also communicate instance-wise uncertainty, i.e., adaptation. To further understand why retaining MSP in SAPS can decrease the set size of easy examples, we present the ground-truth label ranking of different MSP for three different models on ImageNet in Figure 3a. We can observe that MSP is negatively related to ground-truth label ranking. Therefore, the ground-truth labels of easy examples (high MSP values) are generally highly ranked, which are expected to give compact prediction sets. In other words, the maximum softmax probability of a sample contains the information of its difficulty, allowing SAPS to produce adaptive prediction sets.

Here, we further discuss the effect of other softmax probability values on set size. Specifically, we retain the first k maximum softmax probability values, where $k \in \{1, 2, 3, 4, 5\}$. We conduct experiments on the ImageNet with four classifiers of different architectures. In Figure 3a, we show that retaining MSP achieves the best performance on most models. Additionally, more unreliable probability values involved in the score function will lead to larger prediction sets. Overall, retaining MSP in SAPS is the key to decreasing set size and preserving uncertainty.

To explore the sufficiency of MSP for conformal prediction, we provide an ablation study with some variants of SAPS by maintaining more probability values. Specifically, we compare five variants that preserve the top $k \in \{1, 2, 3, 4, 5\}$ values of softmax probabilities, respectively. The results of

ImageNet using the four classifiers are reported in Table 4 of the manuscript. The results show that the vanilla SAPS with only MSP can achieve the best performance in most cases. For instance, with the ResNeXt101 model, only keeping the maximum softmax probability boosts the set size of $k = 5$ from 2.42 to 1.80, a 0.62 of direct improvement. In addition, retaining more tail-end probability values in the score function leads to larger prediction sets. Overall, retaining MSP is sufficient for SAPS.

Effect of λ and the calibration size. In SAPS, we choose an optimal λ by performing a search over a sequence to minimize the set size on a validation set. In this work, the validation set constitutes 20% of the calibration set. Here, we provide an experimental procedure to show *whether set size is sensitive to λ and calibration size.* To this end, we conduct two experiments on ImageNet to analyze the effects of λ and the size of the calibration set.

We present the results of four models in Figure 3. Indeed, Figure 3b illustrates that one can efficiently utilize grid search to find the optimal λ. Furthermore, as depicted in Figure 3c, nearly all models maintain stable results when the number of calibration sets increases. Overall, the results demonstrate that the set size is not sensitive to variations in λ and the calibration size.

SAPS vs. RAPS ($k_r = 1$). While SAPS has demonstrated strong promise, it shares a similarity in the definition of non-conformity scores with RAPS ($k_r = 1$), as shown below:

$$S_{raps}(\boldsymbol{x}, y, u, \hat{\pi}) = \sum_{i=1}^{o(y,\hat{\pi}(\boldsymbol{x}))-1} \hat{\pi}_{(i)}(\boldsymbol{x}) + u * \hat{\pi}_{o(y,\hat{\pi}(\boldsymbol{x}))}(\boldsymbol{x}) + \phi \cdot (o(y, \hat{\pi}(\boldsymbol{x})) - k_r)^+.$$

Here, ϕ represents the weight of regularization and $(z)^+$ denotes the positive part of z. To this end, we conduct experiments on CIFAR-10, CIFAR-100, and ImageNet to compare SAPS and RAPS ($k_r = 1$), with $\alpha = 0.1$.

As indicated in Table 3, SAPS outperforms RAPS ($k_r = 1$) in large-scale classification scenarios, achieving smaller prediction sets and lower conditional coverage violations. In the small-scale classification task (i.e., CIFAR-10), SAPS produces a comparable set size with RAPS ($k_r = 1$), and the ESCV of SAPS was more than 1.5 times as small as those from RAPS. Overall, employing a constant to substitute the noisy probabilities is an effective way to alleviate the negative implications of noisy probabilities further.

Relation to temperature scaling. In the literature, temperature scaling calibrates softmax probabilities output of models by minimizing Expected Calibration Error (ECE), leading to a reliable maximum probability (Guo et al., 2017). Moreover, as defined in Equation 10, the remaining probability value used in the non-conformity score function is the maximum probability. Thus, a question arises: *what is a relation between temperature scaling and the set size of SAPS?* Here, we vary the value of temperature $T = \{0.1, 0.5, 1, 1.1, 1.3, 1.5, 1.7, 1.9, 2, 5, 10, 20\}$ in temperature scaling. We utilize SAPS to test the ResNet152 model calibrated by different temperatures on the ImageNet benchmark. The results indicate that there exists a consistency between the temperature value and the set size.

As illustrated in Figure 3d, the optimal temperature on ECE, i.e., 1.3coincides with the temperature that yields the smallest set size. Moreover, we can observe that an increase in ECE results in an expanded set size. Indeed, temperature scaling can not change the permutation of the softmax probabilities but improves the confidence level of the maximum probability, resulting in the non-conformity scores of SAPS being more reliable. Overall, for SAPS, confidence calibration is important to generate compact prediction sets.

6. Related work

Conformal prediction (CP) is a statistical framework characterized by a finite-sample coverage guarantee (Vovk et al., 2005). It has been utilized in various tasks including regression (Lei & Wasserman, 2014; Romano et al., 2019), classification (Sadinle et al., 2019), structured prediction (Bates et al., 2021), Large-Language Model (Kumar et al., 2023; Ren et al., 2023; Quach et al., 2023; Su et al., 2024), Diffusion Model (Teneggi et al., 2023; Horwitz & Hoshen, 2022), robots control (Wang et al., 2023), graph neural network (Wijegunawardana et al., 2020; Clarkson, 2023; Song et al., 2024) and so on. Conformal prediction also is deployed in other applications, such as human-in-the-loop decision making (Straitouri et al., 2023; Cresswell et al., 2024), automated vehicles (Bang et al., 2024), and scientific machine learning (Moya et al., 2024; Podina et al., 2024). Besides the inductive conformal prediction (Vovk et al., 2005), most CP methods follow an inductive conformal framework (or split conformal prediction) with a held-out calibration set to cope with the relative computational inefficiency of conformal predictors (Papadopoulos et al., 2002a;b; Vovk, 2002), as in the main paper (Lei et al., 2015). Additionally, there are various variants of conformal predictors based on cross-validation (Vovk, 2015) or jackknife (i.e., leave-one-out) (Barber et al., 2021). The primary focal points of CP are reducing prediction set size and enhancing coverage rate.

Inefficiency. Strategies to improve the efficiency of prediction sets can be roughly split into the following two branches. The first avenue of research focuses on develop-

ing new training algorithms to reduce the average prediction set size (Bellotti, 2021; Colombo & Vovk, 2020; Chen et al., 2021; Stutz et al., 2022; Einbinder et al., 2022; Bai et al., 2022; Fisch et al., 2021; Yang & Kuchibhotla, 2021). Those training methods are usually computationally expensive due to the model retraining and the high complexity of optimization. Thus, CP has also been studied in the context of Early Stopping (Liang et al., 2023), allowing model selection leading to compact prediction sets while keeping coverage guarantees. The other avenue involves leveraging post-hoc technologies, such as novel score functions (Romano et al., 2020; Angelopoulos et al., 2021; Ghosh et al., 2023) and post-hoc learning method (Xi et al., 2024). There exist others concentrating on unique settings such as federated learning (Lu et al., 2023; Plassier et al., 2023), multi-label problem (Cauchois et al., 2021; Fisch et al., 2022; Papadopoulos, 2014), outlier detection (Bates et al., 2023; Guan & Tibshirani, 2022), and out-of-distribution detection (Chen et al., 2023; Novello et al., 2024). Most existing post-hoc methods calculate the non-conformity score based on unreliable probability values, leading to sub-optimal performance. In contrast to previous post-hoc methods, we show that probability value is not necessary for non-conformity scores and design an effective method to remove the probability value while retaining uncertainty information.

Validity of coverage rate. There is an increasing amount of research focused on improving coverage rates(Shi et al., 2013; Löfström et al., 2015; Ding et al., 2023), including efforts to maintain the marginal coverage rate by modifying the assumption of exchangeability to accommodate factors such as adversaries (Gendler et al., 2021; Kang et al., 2024), covariate shifts (Tibshirani et al., 2019; Deng et al., 2023; Xu et al., 2024), label shifts (Podkopaev & Ramdas, 2021; Plassier et al., 2023) and noisy labels (Feldman et al., 2023; Sesia et al., 2023). Moreover, conformal prediction algorithms pursue conditional coverage guarantee (Vovk, 2012), including training-conditional validity and object-conditional validity. Specifically, some methods acquire the training-conditional coverage guarantee which would ensure that most draws of the training data set to result in valid marginal coverage on future test points (Bian & Barber, 2023; Pournaderi & Xiang, 2024). There are also group-conditional conformal prediction methods that guarantee performance on all groups of individuals in the population (Javanmard et al., 2022; Gibbs et al., 2023; Melki et al., 2023). Although it is impossible to obtain conditional coverage (Foygel Barber et al., 2021), THR (Sadinle et al., 2019) shows that efficient class-conditional coverage is possible. For example, Clustered CP (Ding et al., 2023) utilizes the taxonomy of the label space to improve the class-conditional coverage rate and k-Class-conditional CP calibrates the class-specific scores threshold depending on the top-k error. As discussed in Sec.4.2, SAPS broadly enhances the conditional coverage rate of prediction sets while maintaining small prediction sets.

7. Conclusion

In this paper, we present SAPS, a simple and effective conformal prediction score function that discards almost all the probability values except for the maximum softmax probability. By highlighting the information on label ranking, SAPS effectively mitigates the negative effect of tail probabilities, resulting in small and stable prediction sets. Additionally, we provide a key insight that the value of softmax probability might be redundant information for conformal prediction. Extensive experiments show that SAPS can improve the conditional coverage rate while maintaining a small prediction set. Our method is straightforward to implement with any classifier and can be easily adopted in various practical settings. We expect that out method can inspire future research to emphasize the label ranking information for conformal prediction.

8. Limitations

The limitation of our method is mainly on the hyperparameter representing the weight of label ranking that requires to be tuned with a validation dataset (same as RAPS). About future direction, we expect to enhance the conditional coverage of these conformal prediction algorithms on large-scale datasets, where all CP methods achieve worse conditional coverage rates than those of small-scale datasets.

Acknowledgements

This research is supported by the Shenzhen Fundamental Research Program (Grant No. JCYJ20230807091809020). Yue Qiu is supported by the National Natural Science Foundation of China (Grant No. 12101407), the Chongqing Entrepreneurship and Innovation Program for Returned Overseas Scholars (Grant No. CX2023068), and the Fundamental Research Funds for the Central Universities (Grant No. 2022CDJJCLK-002, 2023CDJXY-042). We gratefully acknowledge the support of the Center for Computational Science and Engineering at the Southern University of Science and Technology, as well as the Key Laboratory of Nonlinear Analysis and its Applications at Chongqing University, under the Ministry of Education, China, for our research.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

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

Angelopoulos, A. N., Bates, S., Jordan, M. I., and Malik, J. Uncertainty sets for image classifiers using conformal prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Bai, Y., Mei, S., Wang, H., Zhou, Y., and Xiong, C. Efficient and differentiable conformal prediction with general function classes. *arXiv preprint arXiv:2202.11091*, 2022.

Balasubramanian, V., Ho, S.-S., and Vovk, V. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

Bang, H., Dave, A., and Malikopoulos, A. A. Safe merging in mixed traffic with confidence. *arXiv preprint arXiv:2403.05742*, 2024.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486 – 507, 2021.

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

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023.

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

Bian, M. and Barber, R. F. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, pp. 2044–2066, 2023.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pp. 1721–1730, 2015.

Cauchois, M., Gupta, S., and Duchi, J. C. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *The Journal of Machine Learning Research*, 22(1):3681–3722, 2021.

Chen, H., Huang, Z., Lam, H., Qian, H., and Zhang, H. Learning prediction intervals for regression: Generalization and calibration. In *International Conference on Artificial Intelligence and Statistics*, pp. 820–828. PMLR, 2021.

Chen, X., Li, Y., and Yang, Y. Batch-ensemble stochastic neural networks for out-of-distribution detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Clarkson, J. Distribution free prediction sets for node classification. In *International Conference on Machine Learning*, pp. 6268–6278. PMLR, 2023.

Colombo, N. and Vovk, V. Training conformal predictors. In *Conformal and Probabilistic Prediction and Applications*, pp. 55–64. PMLR, 2020.

Cresswell, J. C., Sui, Y., Kumar, B., and Vouitsis, N. Conformal prediction sets improve human decision making. *arXiv preprint arXiv:2401.13744*, 2024.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Deng, Z., Dwork, C., and Zhang, L. Happymap: A generalized multicalibration method. In *14th Innovations in Theoretical Computer Science Conference, ITCS 2023*, pp. 41. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2023.

Ding, T., Angelopoulos, A. N., Bates, S., Jordan, M. I., and Tibshirani, R. J. Class-conditional conformal prediction with many classes. *arXiv preprint arXiv:2306.09335*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

Feldman, S., Einbinder, B.-S., Bates, S., Angelopoulos, A. N., Gendler, A., and Romano, Y. Conformal prediction is robust to dispersive label noise. In *Conformal and Probabilistic Prediction with Applications*, pp. 624–626. PMLR, 2023.

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

Fisch, A., Schuster, T., Jaakkola, T., and Barzilay, R. Conformal prediction sets with limited false positives. In *International Conference on Machine Learning*, pp. 6514–6532. PMLR, 2022.

Foygel Barber, R., Candes, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482, 2021.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059. PMLR, 2016.

Gendler, A., Weng, T.-W., Daniel, L., and Romano, Y. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2021.

Ghosh, S., Belkhouja, T., Yan, Y., and Doppa, J. R. Improving uncertainty quantification of deep classifiers via neighborhood conformal prediction: Novel algorithm and theoretical analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7722–7730. AAAI Press, 2023.

Gibbs, I., Cherian, J. J., and Candès, E. J. Conformal prediction with conditional guarantees. *arXiv preprint arXiv:2305.12616*, 2023.

Guan, L. and Tibshirani, R. Prediction and outlier detection in classification problems. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):524–546, 2022.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2016.

Horwitz, E. and Hoshen, Y. Conffusion: Confidence intervals for diffusion models. *arXiv preprint arXiv:2211.09795*, 2022.

Javanmard, A., Shao, S., and Bien, J. Prediction sets for high-dimensional mixture of experts models. *arXiv preprint arXiv:2210.16710*, 2022.

Kang, M., Gürel, N. M., Li, L., and Li, B. Colep: Certifiably robust learning-reasoning conformal prediction via probabilistic circuits. *arXiv preprint arXiv:2403.11348*, 2024.

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

Kumar, B., Lu, C., Gupta, G., Palepu, A., Bellamy, D., Raskar, R., and Beam, A. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

Lei, J. and Wasserman, L. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1): 71–96, 2014.

Lei, J., Rinaldo, A., and Wasserman, L. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74:29–43, 2015.

Liang, Z., Zhou, Y., and Sesia, M. Conformal inference is (almost) free for neural networks trained with early stopping. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20810–20851. PMLR, 2023.

Löfström, T., Boström, H., Linusson, H., and Johansson, U. Bias reduction through conditional conformal prediction. *Intelligent Data Analysis*, 19(6):1355–1375, 2015.

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

Melki, P., Bombrun, L., Diallo, B., Dias, J., and Da Costa, J.-P. Group-conditional conformal prediction via quantile regression calibration for crop and weed classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 614–623, October 2023.

Moya, C., Mollaali, A., Zhang, Z., Lu, L., and Lin, G. Conformalized-deeponet: A distribution-free framework for uncertainty quantification in deep operator networks. *arXiv preprint arXiv:2402.15406*, 2024.

Novello, P., Dalmau, J., and Andeol, L. Out-of-distribution detection should use conformal prediction (and vice-versa?). *arXiv preprint arXiv:2403.11532*, 2024.

Papadopoulos, H. A cross-conformal predictor for multi-label classification. In *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings 10*, pp. 241–250. Springer, 2014.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, pp. 345–356. Springer Berlin Heidelberg, 2002a.

Papadopoulos, H., Vovk, V., and Gammerman, A. Qualified prediction for large data sets in the case of pattern recognition. In *International Conference on Machine Learning and Applications*, 2002b.

Plassier, V., Makni, M., Rubashevskii, A., Moulines, E., and Panov, M. Conformal prediction for federated uncertainty quantification under label shift. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27907–27947. PMLR, 2023.

Podina, L., Rad, M. T., and Kohandel, M. Conformalized physics-informed neural networks. *arXiv preprint arXiv:2405.08111*, 2024.

Podkopaev, A. and Ramdas, A. Distribution-free uncertainty quantification for classification under label shift. In *Uncertainty in Artificial Intelligence*, pp. 844–853. PMLR, 2021.

Pournaderi, M. and Xiang, Y. Training-conditional coverage bounds under covariate shift, 2024.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Ren, A. Z., Dixit, A., Bodrova, A., Singh, S., Tu, S., Brown, N., Xu, P., Takayama, L., Xia, F., Varley, J., et al. Robots that ask for help: Uncertainty alignment for large language model planners. In *Conference on Robot Learning*, pp. 661–682. PMLR, 2023.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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

Sesia, M., Wang, Y. R., and Tong, X. Adaptive conformal classification with noisy labels. In *2023 IMS International Conference on Statistics and Data Science (ICSDS)*, pp. 221, 2023.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Shi, F., Ong, C. S., and Leckie, C. Applications of class-conditional conformal predictor in multi-class classification. In *2013 12th International Conference on Machine Learning and Applications*, volume 1, pp. 235–239, 2013. doi: 10.1109/ICMLA.2013.48.

Smith, R. C. *Uncertainty quantification: theory, implementation, and applications*, volume 12. Siam, 2013.

Song, J., Huang, J., Jiang, W., Zhang, B., Li, S., and Wang, C. Similarity-navigated conformal prediction for graph neural networks. *arXiv preprint arXiv:2405.14303*, 2024.

Straitouri, E., Wang, L., Okati, N., and Rodriguez, M. G. Improving expert predictions with conformal prediction. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Stutz, D., Dvijotham, K., Cemgil, A. T., and Doucet, A. Learning optimal conformal classifiers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

Su, J., Luo, J., Wang, H., and Cheng, L. Api is enough: Conformal prediction for large language models without logit-access. *arXiv preprint arXiv:2403.01216*, 2024.

Teneggi, J., Tivnan, M., Stayman, W., and Sulam, J. How to trust your diffusion model: A convex optimization approach to conformal risk control. In *International Conference on Machine Learning*, pp. 33940–33960. PMLR, 2023.

Teng, J., Wen, C., Zhang, D., Bengio, Y., Gao, Y., and Yuan, Y. Predictive inference with feature conformal prediction. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*, 2023.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Vovk, V. On-line confidence machines are well-calibrated. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pp. 187–196, 2002.

Vovk, V. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*, pp. 475–490. PMLR, 2012.

Vovk, V. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28, 2015.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Wang, J., Tong, J., Tan, K., Vorobeychik, Y., and Kantaros, Y. Conformal temporal logic planning using large language models: Knowing when to do what and when to ask for help. *arXiv preprint arXiv:2309.10092*, 2023.

Wijegunawardana, P., Gera, R., and Soundarajan, S. Node classification with bounded error rates. In *Complex Networks XI: Proceedings of the 11th Conference on Complex Networks CompleNet 2020*, pp. 26–38. Springer, 2020.

Xi, H., Huang, J., Feng, L., and Wei, H. Does confidence calibration help conformal prediction? *arXiv preprint arXiv:2402.04344*, 2024.

Xu, R., Sun, Y., Chen, C., Venkitasubramaniam, P., and Xie, S. Robust conformal prediction under distribution shift via physics-informed structural causal model. *arXiv preprint arXiv:2403.15025*, 2024.

Yang, Y. and Kuchibhotla, A. K. Finite-sample efficient conformal prediction. *arXiv preprint arXiv:2104.13871*, 2021.

A. Proof of Proposition 3.2

Proof. Let $\hat{\pi}$ be a classifier. $(\boldsymbol{x}_1, y_1)$ and $(\boldsymbol{x}_2, y_2)$ represent two instances sampled from the joint data distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$, and u_1, u_2 are sampled from $U[0,1]$. It is easy to get that for any different positive numbers γ_1 and γ_2, $\gamma_1 \cdot [o(y_1, \hat{\pi}(\boldsymbol{x}_1)) - 1 + u_1] \leq \gamma_1 \cdot [o(y_2, \hat{\pi}(\boldsymbol{x}_2)) - 1 + u_2]$ if and only if $\gamma_2 \cdot [o(y_1, \hat{\pi}(\boldsymbol{x}_1)) - 1 + u_1] \leq \gamma_2 \cdot [o(y_2, \hat{\pi}(\boldsymbol{x}_2)) - 1 + u_2]$. According to Lemma 3.1, we can get that $S_{cons}(\boldsymbol{x}, y; \hat{\pi}, \gamma_1)$ and $S_{cons}(\boldsymbol{x}, y; \hat{\pi}, \gamma_2)$ could provide the same prediction sets. □

B. Proof of Theorem 3.3

Proof. For simplicity, let $a_r = A_r - A_{r-1}$ for $r \geq 1$ and $A_0 = 0$. a_r represents the rate of examples in the calibration set whose ground-truth labels' rank is r. The non-conformity scores of examples whose rank of ground-truth labels are r can be computed by

$$S_{cons}(\boldsymbol{x}, y, u; \hat{\pi}) = r - u,$$

where $u \sim U[0,1]$. Thus, we can observe that the scores of these examples are uniformly distributed on $[r-1, r]$. Scores of examples whose rank of ground-truth labels are less than or equal to r are no more than r. In other words, the rate of examples in the calibration set with scores lower than r is A_r. Given that k satisfies $A_k \geq 1-\alpha > A_{k-1}$, the $1-\alpha$ quantile of scores for the calibration set, i.e., the calibrated threshold τ, is located in the interval $[k-1, k]$. Since the scores in the interval $[k-1, k]$ are uniformly distributed and the rate of examples whose scores in $[k-1, k]$ is a_k, τ is equivalent to the $\frac{1-\alpha-A_{k-1}}{a_k}$ quantile of $[k-1, k]$. Formally, we can get the value of τ by

$$\tau = (k-1) + \frac{1-\alpha-A_{k-1}}{a_k}.$$

Based on the definition of prediction set in Equation 3, the prediction set of a test instance $\boldsymbol{x}$ is equal to

$$\mathcal{C}_{1-\alpha}(\boldsymbol{x}, u) = \{y \in \mathcal{Y} : o(y, \hat{\pi}(\boldsymbol{x})) - u \leq \tau\}. \tag{11}$$

Next, we analyze the prediction set $\mathcal{C}_{1-\alpha}(\boldsymbol{x}, u)$. When $y \in \mathcal{Y}$ satisfies the inequation $o(y, \hat{\pi}(\boldsymbol{x})) \leq k-1$, $S_{cons}(\boldsymbol{x}, y, u; \hat{\pi})$ must be smaller than τ. Thus, the size of sets is at least $k-1$. But for y satisfying $o(y, \hat{\pi}(\boldsymbol{x})) = k$, we can have

$$\begin{aligned} o(y, \hat{\pi}(\boldsymbol{x})) - u &\leq \tau \\ k - u &\leq (k-1) + \frac{1-\alpha-A_{k-1}}{a_k} \\ 1 - \frac{1-\alpha-A_{k-1}}{a_k} &\leq u \end{aligned}$$

Thus, there exists a probability of $\frac{1-\alpha-A_{k-1}}{a_k}$ such that $S(\boldsymbol{x}, y, u; \hat{\pi}) \leq \tau$. Finally, the expected value of the set size for the test example $\boldsymbol{x}$ is

$$\mathbb{E}_{u \sim U}[|\mathcal{C}_{1-\alpha}(\boldsymbol{x}, u)|] = k - 1 + \frac{1-\alpha-A_{k-1}}{a_k} = k - 1 + \frac{1-\alpha-A_{k-1}}{A_k - A_{k-1}} \tag{12}$$

□

C. Detailed results for coverage rate and set size

In this section, we report the detailed results of coverage rate and set size on different datasets in Table 5,6, and 7. The median-of-means for each result is reported over ten different trials. The average results of multiple models have been reported in the main paper.

Table 5. Results on ImageNet. The median-of-means for each column is reported over 10 different trials. **Bold** numbers indicate optimal performance.

	$\alpha = 0.1$						$\alpha = 0.05$					
	Coverage			Size ↓			Coverage			Size ↓		
Datasets	APS	RAPS	SAPS	APS	RAPS	SAPS	APS	RAPS	SAPS	APS	RAPS	SAPS
ResNeXt101	0.899	0.902	0.901	19.49	2.01	**1.82**	0.950	0.951	0.950	46.58	4.24	**3.83**
ResNet152	0.900	0.900	0.900	10.51	2.10	**1.92**	0.950	0.950	0.950	22.65	4.39	**4.07**
ResNet101	0.898	0.900	0.900	10.83	2.24	**2.07**	0.948	0.949	0.950	23.20	4.78	**4.34**
ResNet50	0.899	0.900	0.900	12.29	2.51	**2.31**	0.948	0.950	0.950	25.99	5.57	**5.25**
ResNet18	0.899	0.900	0.900	16.10	4.43	**4.00**	0.949	0.950	0.950	32.89	11.75	**10.47**
DenseNet161	0.900	0.900	0.900	12.03	2.27	**2.08**	0.949	0.950	0.951	28.06	5.11	**4.61**
VGG16	0.897	0.901	0.900	14.00	3.59	**3.25**	0.948	0.950	0.949	27.55	8.80	**7.84**
Inception	0.900	0.902	0.902	87.93	5.32	**4.58**	0.949	0.951	0.950	167.98	18.71	**14.43**
ShuffleNet	0.900	0.899	0.900	31.77	5.04	**4.54**	0.949	0.950	0.950	69.39	16.13	**14.05**
ViT	0.900	0.898	0.900	10.55	1.70	**1.61**	0.950	0.949	0.950	31.75	3.91	**3.21**
DeiT	0.901	0.900	0.900	8.51	1.48	**1.41**	0.950	0.949	0.949	24.88	2.69	**2.49**
CLIP	0.899	0.900	0.900	17.45	6.81	**6.23**	0.951	0.949	0.949	35.09	16.79	**16.07**
average	0.899	0.900	0.900	20.95	3.29	**2.98**	0.949	0.950	0.950	44.67	8.57	**7.55**

Table 6. Results on CIFAR-100. The median-of-means for each column is reported over 10 different trials. **Bold** numbers indicate optimal performance.

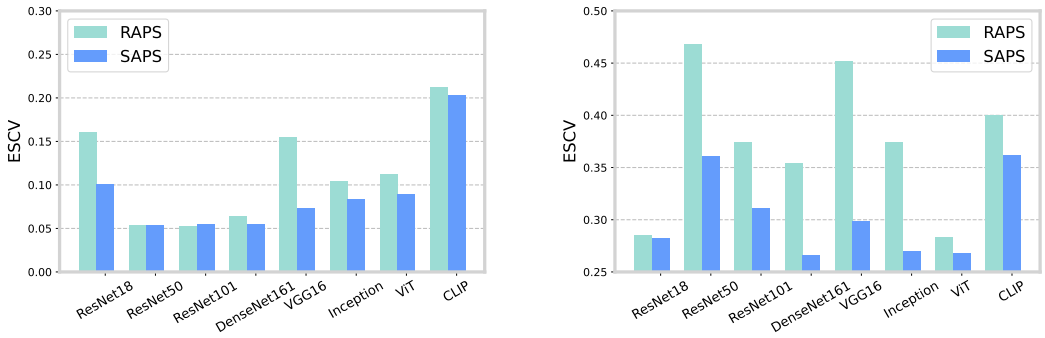
	$\alpha = 0.1$						$\alpha = 0.05$					
	Coverage			Size ↓			Coverage			Size ↓		
Datasets	APS	RAPS	SAPS	APS	RAPS	SAPS	APS	RAPS	SAPS	APS	RAPS	SAPS
ResNet18	0.898	0.901	0.898	10.03	2.72	**2.41**	0.949	0.950	0.951	16.76	5.88	**4.96**
ResNet50	0.896	0.902	0.899	6.51	2.16	**2.04**	0.946	0.948	0.946	12.49	4.37	**3.69**
ResNet101	0.899	0.901	0.898	6.52	2.10	**1.99**	0.951	0.947	0.948	12.26	4.49	**3.85**
DenseNet161	0.898	0.898	0.897	8.07	2.02	**1.77**	0.948	0.949	0.950	14.35	3.61	**3.33**
VGG16	0.900	0.896	0.895	5.80	4.58	**3.64**	0.951	0.949	0.948	11.83	11.32	**9.27**
Inception	0.902	0.902	0.902	12.01	2.01	**2.01**	0.952	0.953	0.953	18.24	5.21	**4.26**
ViT	0.897	0.901	0.900	4.29	2.14	**1.91**	0.949	0.948	0.949	7.92	3.98	**3.57**
CLIP	0.899	0.900	0.900	9.84	6.18	**5.58**	0.952	0.948	0.950	16.04	12.50	**11.27**
average	0.899	0.900	0.899	7.88	2.99	**2.67**	0.950	0.949	0.949	13.74	6.42	**5.53**

Table 7. Results on CIFAR-10. The median-of-means for each column is reported over 10 different trials. **Bold** numbers indicate optimal performance.

	$\alpha = 0.1$						$\alpha = 0.05$					
	Coverage			Size ↓			Coverage			Size ↓		
Datasets	APS	RAPS	SAPS	APS	RAPS	SAPS	APS	RAPS	SAPS	APS	RAPS	SAPS
ResNet18	0.896	0.896	0.898	2.42	2.29	**2.18**	0.948	0.947	0.949	3.12	3.12	**3.10**
ResNet50	0.897	0.900	0.895	2.08	1.95	**1.77**	0.949	0.950	0.950	2.69	2.62	**2.53**
ResNet101	0.899	0.901	0.898	2.15	2.00	**1.86**	0.949	0.948	0.949	2.78	2.68	**2.59**
DenseNet161	0.898	0.899	0.898	2.06	1.90	**1.71**	0.949	0.953	0.949	2.67	2.53	**2.35**
VGG16	0.897	0.900	0.897	1.75	1.52	**1.39**	0.950	0.949	0.949	2.22	2.02	**1.87**
Inception	0.904	0.904	0.902	2.28	2.04	**1.78**	0.952	0.953	0.952	3.04	2.78	**2.55**
ViT	0.901	0.899	0.900	1.50	1.34	**1.21**	0.952	0.952	0.950	1.89	1.79	**1.58**
CLIP	0.897	0.900	0.899	1.51	1.27	**1.16**	0.949	0.949	0.950	1.89	1.57	**1.41**
average	0.899	0.900	0.898	1.97	1.79	**1.63**	0.950	0.950	0.950	2.54	2.39	**2.25**

D. ESCV on CIAFR-10 and CIFAR-100

In this section, we report the results of ESCV on CIFAR-10 and CIFAR-100 with $\alpha = 0.1$. From Figure 4a and 4b, we can observe that the SAPS still can get smaller ESCV than RAPS. Moreover, as the scale of the classification task increases, the ESCV for RAPS also exhibits a marked increase, whereas SAPS consistently demonstrates a low value. For example, with the CLIP model, the ESCV for RAPS on ImageNet exceeds that observed for CIFAR-100. But, for both ImageNet and CIFAR-100, the ESCV of SAPS remains around 0.35.



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

Figure 4. ESCV for different models on CIFAR-10 and CIFAR-100 with $\alpha = 0.1$.

E. SAPS vs. RAPS($k_r = 1$)

To further understand the influence of excluding the probability value, we conduct an experiment for RAPS($k_r = 1$), which is similar to SAPS on ImageNet, CIFAR-100, and CIFAR-10 with $\alpha = 0.1$. The median results on ten trials are reported in Table 8, 9 and 10.

Table 8. Comparsion between RAPS($k_r = 1$) and SAPS on ImageNet with $\alpha = 0.1$. The median-of-means for each column is reported over 10 different trials. **Bold** numbers indicate optimal performance.

	$\alpha = 0.1$						$\alpha = 0.05$					
	Coverage		Size ↓		ESCV ↓		Coverage		Size ↓		ESCV ↓	
Datasets	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS
ResNeXt101	0.902	0.901	1.85	**1.82**	0.477	**0.366**	0.950	0.950	4.56	**3.83**	0.505	**0.358**
ResNet152	0.901	0.900	1.97	**1.92**	0.572	**0.440**	0.950	0.950	4.68	**4.07**	0.750	**0.335**
ResNet101	0.900	0.900	2.09	**2.07**	0.650	**0.431**	0.950	0.950	4.95	**4.34**	0.724	**0.466**
ResNet50	0.900	0.900	2.32	**2.31**	0.594	**0.350**	0.950	0.950	5.52	**5.25**	0.950	**0.373**
ResNet18	0.899	0.900	4.50	**4.00**	0.900	**0.409**	0.950	0.950	**10.03**	10.47	0.950	**0.490**
DenseNet161	0.901	0.900	2.12	**2.08**	0.531	**0.393**	0.951	0.951	5.15	**4.61**	0.950	**0.405**
VGG16	0.899	0.900	3.69	**3.25**	0.671	**0.425**	0.949	0.949	**7.56**	7.84	0.950	**0.390**
Inception	0.901	0.902	5.13	**4.58**	0.478	**0.418**	0.951	0.950	**13.20**	14.43	0.950	**0.532**
ShuffleNet	0.900	0.900	5.11	**4.54**	0.481	**0.375**	0.951	0.950	**12.38**	14.05	0.473	**0.306**
ViT	0.899	0.900	**1.58**	1.61	0.900	**0.393**	0.950	0.950	3.32	**3.21**	0.950	**0.950**
DeiT	0.899	0.900	**1.38**	1.41	0.416	**0.389**	0.950	0.949	2.59	**2.49**	0.596	**0.418**
CLIP	0.900	0.900	7.16	**6.23**	0.900	**0.369**	0.949	0.949	**15.16**	16.07	0.950	**0.317**
average	0.900	0.900	3.24	**2.98**	0.631	**0.396**	0.950	0.950	**7.43**	7.55	0.808	**0.445**

Table 9. Comparsion between RAPS($k_r = 1$) and SAPS on CIFAR-100 with $\alpha = 0.1$. The median-of-means for each column is reported over 10 different trials. **Bold** numbers indicate optimal performance.

	$\alpha = 0.1$						$\alpha = 0.05$					
	Coverage		Size ↓		ESCV ↓		Coverage		Size ↓		ESCV ↓	
Datasets	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS
ResNet18	0.897	0.898	2.51	**2.41**	0.321	**0.282**	0.951	0.951	6.30	**4.96**	**0.209**	0.244
ResNet50	0.900	0.899	**1.97**	2.04	0.383	**0.360**	0.948	0.946	4.52	**3.69**	0.396	**0.200**
ResNet101	0.899	0.898	**1.92**	1.99	0.344	**0.311**	0.950	0.948	4.56	**3.85**	0.210	**0.242**
DenseNet161	0.899	0.897	**1.75**	1.77	0.381	**0.265**	0.949	0.950	3.69	**3.33**	0.303	**0.298**
VGG16	0.897	0.895	4.06	**3.64**	0.567	**0.298**	0.950	0.948	10.75	**9.27**	0.450	**0.325**
Inception	0.901	0.902	**1.86**	2.01	0.375	**0.269**	0.953	0.953	4.57	**4.26**	0.305	**0.213**
ViT	0.902	0.900	**1.89**	1.91	0.312	**0.268**	0.946	0.949	4.01	**3.57**	0.190	**0.170**
CLIP	0.900	0.900	6.38	**5.58**	0.438	**0.362**	0.949	0.950	**10.47**	11.27	**0.406**	0.429
average	0.899	0.899	2.79	**2.67**	0.390	**0.302**	0.949	0.949	6.11	**5.53**	0.309	**0.265**

Table 10. Comparsion between RAPS($k_r = 1$) and SAPS on CIFAR-10 with $\alpha = 0.1$. The median-of-means for each column is reported over 10 different trials. **Bold** numbers indicate optimal performance.

	$\alpha = 0.1$						$\alpha = 0.05$					
	Coverage		Size ↓		ESCV ↓		Coverage		Size ↓		ESCV ↓	
Datasets	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS	RAPS ($k_r = 1$)	SAPS
ResNet18	0.897	0.898	2.24	**2.18**	0.145	**0.100**	0.946	0.949	**2.98**	3.10	**0.052**	0.076
ResNet50	0.899	0.895	1.79	**1.77**	0.101	**0.054**	0.949	0.950	2.63	**2.53**	0.062	**0.030**
ResNet101	0.899	0.898	1.87	**1.86**	0.107	**0.054**	0.946	0.949	2.61	**2.59**	0.043	**0.026**
DenseNet161	0.899	0.898	**1.69**	1.71	0.156	**0.055**	0.952	0.949	2.52	**2.35**	0.103	**0.021**
VGG16	0.900	0.897	**1.38**	1.39	0.112	**0.073**	0.949	0.949	**1.84**	1.87	0.124	**0.050**
Inception	0.903	0.902	**1.77**	1.78	0.143	**0.083**	0.954	0.952	2.68	**2.55**	0.089	**0.040**
ViT	0.900	0.900	**1.16**	1.21	0.127	**0.089**	0.950	0.950	1.63	**1.58**	0.101	**0.054**
CLIP	0.899	0.899	**1.06**	1.16	0.213	**0.203**	0.950	0.950	**1.36**	1.41	0.254	**0.123**
average	0.900	0.898	**1.62**	1.63	0.138	**0.089**	0.949	0.950	2.28	**2.25**	0.104	**0.052**