Improving Uncertainty Quantification of Deep Classifiers via Neighborhood Conformal Prediction: Novel Algorithm and Theoretical Analysis

Subhankar Ghosh*, Taha Belkhouja*, Yan Yan, Janardhan Rao Doppa

School of EECS, Washington State University
{subhankar.ghosh, taha.belkhouja, yan.yan1, jana.doppa}@wsu.edu

Abstract

Safe deployment of deep neural networks in high-stake real-world applications requires theoretically sound uncertainty quantification. Conformal prediction (CP) is a principled framework for uncertainty quantification of deep models in the form of prediction set for classification tasks with a user-specified coverage (i.e., true class label is contained with high probability). This paper proposes a novel algorithm referred to as *Neighborhood Conformal Prediction (NCP)* to improve the efficiency of uncertainty quantification from CP for deep classifiers (i.e., reduce prediction set size). The key idea behind NCP is to use the learned representation of the neural network to identify k nearest-neighbors calibration examples for a given testing input and assign them importance weights proportional to their distance to create adaptive prediction sets. We theoretically show that if the learned data representation of the neural network satisfies some mild conditions, NCP will produce smaller prediction sets than traditional CP algorithms. Our comprehensive experiments on CIFAR-10, CIFAR-100, and ImageNet datasets using diverse deep neural networks strongly demonstrate that NCP leads to significant reduction in prediction set size over prior CP methods.

1 Introduction

Recent advances in deep learning have allowed us to build models with high accuracy. However, to safely deploy these deep models in high-stake applications (e.g, medical diagnosis) for critical decision-making, we need theoretically-sound uncertainty quantification (UQ) to capture the deviation of the prediction from the ground-truth output. The UQ could take the form of a *prediction set* (a subset of candidate labels) for classification tasks. For example, in medical diagnosis, such prediction sets will allow a doctor to rule out harmful diagnoses such as stomach cancer even if the most likely diagnosis is a stomach ache. Conformal prediction (CP) (Vovk, Gammerman, and Saunders 1999; Vovk, Gammerman, and Shafer 2005; Shafer and Vovk 2008) is a principled framework for UQ that provides formal guarantees for a user-specified *coverage*: ground-truth output is contained in the prediction set with a high probability α (e.g., 90%) for classification. Additionally, UQ from CP is adaptive and will reflect the difficulty of testing inputs: size of the prediction set will be large for difficult inputs and small for easy inputs.

There are two key steps in CP. First, in the prediction step, we use a trained model (e.g., deep neural network) to compute *conformity scores* which measure similarity between calibration examples and a testing input. Second, in the calibration step, we use the conformity scores on a set of calibration examples to find a threshold to construct prediction set which meets the coverage constraint (e.g., α=90%). The *efficiency* of UQ from CP (Sadinle, Lei, and Wasserman 2019) is measured in terms of size of the prediction set (the smaller the better). There is an inherent trade-off between coverage and efficiency. For example, it is easy to achieve high coverage with low efficiency (i.e., large prediction set) by including all or most candidate labels in the prediction set. CP for classification is relatively under-studied. Recent work has proposed conformity scores based on ordered probabilities (Romano, Sesia, and Candes 2020; Angelopoulos et al. 2021) for UQ based on CP for classification and do not come with theoretical guarantees about efficiency. The main research question of this paper is: *how can we improve CP to achieve provably higher efficiency by satisfying the marginal coverage constraint for (pre-trained) deep classifiers?*

To answer this question, this paper proposes a novel algorithm referred to as **N***eighborhood* **C***onformal* **P***rediction (NCP)* that is inspired by the framework of localized CP (LCP) (Guan 2021). The key idea behind LCP is to assign higher importance to calibration examples in the local neighborhood of a given testing input. This is in contrast to the standard CP, which assigns equal importance to all calibration examples. However, there is no theoretical analysis of LCP to characterize the necessary conditions for reduced prediction set size and no empirical evaluation on real-world classification tasks. The effectiveness of LCP critically depends on the localizer and weighting function. The proposed NCP algorithm specifies a concrete localizer and weighting function using the learned input representation from deep neural network classifiers. For a given testing input, NCP identifies k nearest neighbors and assigns them importance weights proportional to their distance defined using the learned representation of deep classifier.

We theoretically analyze the expected threshold of NCP, which is typically used to measure the efficiency of CP algorithms, i.e., a smaller expected threshold indicates smaller

*These authors contributed equally.

prediction sets. Specifically, we prove that if the learned data representation of the neural network satisfies some mild conditions in terms of separation and concentration, NCP will produce smaller prediction sets than traditional CP algorithms. To the best of our knowledge, this is the first result to give affirmative answer to the open question: what are the necessary conditions for NCP-based algorithms to achieve improved efficiency over standard CP? Our theoretical analysis informs us with a principle to design better CP algorithms: automatically train better localizer functions to reduce the prediction set size for classification.

We performed comprehensive experiments on CIFAR-10, CIFAR-100, and ImageNet datasets using a variety of deep neural network classifiers to evaluate the efficacy of NCP and state-of-the-art CP methods. Our results demonstrate that the localizer and weighting function of NCP is highly effective and results in significant reduction in prediction set size; and a better conformity scoring function further improves the efficacy of NCP algorithm. Our ablation experiments clearly match our theoretical analysis of the NCP algorithm.

Contributions. The key contribution of this paper is the development of Neighborhood CP (NCP) algorithm along with its theoretical and empirical analysis for improving the efficiency of uncertainty quantification of deep classifiers.

- Development of the Neighborhood CP algorithm by specifying an effective localizer function using the learned representation from deep classifiers. NCP is complementary to CP methods, i.e., any advancements in CP will automatically benefit NCP for uncertainty quantification.
- Novel theoretical analysis of the Neighborhood CP algorithm to characterize when and why it will improve the efficiency of UQ over standard CP methods.
- Experimental evaluation of NCP on classification benchmarks using diverse deep models to demonstrate its efficacy over prior CP methods. Our code is publicly available on the GitHub repository: https://github.com/1995subhankar1995/NCP.

2 Background and Problem Setup

We consider the problem of uncertainty quantification (UQ) of pre-trained deep models for classification tasks. Suppose x is an input from the space $\mathcal{X}$ and $y^* \in \mathcal{Y}$ is the corresponding ground-truth output. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be the joint space of input-output pairs and the underlying distribution on $\mathcal{Z}$ be $\mathcal{D}_{\mathcal{Z}}$. For classification tasks, $\mathcal{Y}$ is a set of C discrete class-labels $\{1, 2, \cdots, C\}$. As per the standard notation in conformal prediction, X is a random variable and x is a data sample.

Uncertainty quantification. Let $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{cal}}$ correspond to sets of training and calibration examples drawn from a target distribution $\mathcal{D}_{\mathcal{Z}}$. We assume the availability of a pre-trained deep model $F_\theta : \mathcal{X} \mapsto \mathcal{Y}$, where θ stands for the parameters of the deep model. For a given testing input x, we want to compute UQ of the deep model F_θ in the form of a prediction set $\mathcal{C}(x)$, a subset of candidate class-labels $\{1, 2, \cdots, C\}$.

Coverage and efficiency. The performance of UQ is measured using two metrics. First, the (marginal) *coverage* is defined as the probability that the ground-truth output y^* is contained in $\mathcal{C}(x)$ for a testing example (x, y^*) from the same data distribution $\mathcal{D}_{\mathcal{Z}}$, i.e., $\mathbb{P}(y^* \in \mathcal{C}(x))$. The empirical coverage Cov is measured over a given set of testing examples $\mathcal{D}_{\text{test}}$. Second, *efficiency*, denoted by Eff, measures the cardinality of the prediction set $\mathcal{C}(x)$ for classification. Smaller prediction set means higher efficiency. It is easy to achieve the desired coverage (say 90%) by always outputting $\mathcal{C}(x)$=$\mathcal{Y}$ at the expense of poor efficiency.

Conformal prediction (CP). CP is a framework that allows us to compute UQ for any given predictor through a conformalization step. The key element of CP is a measure function V to compute the *conformity* (or *non-conformity*) score, measures similarity between labeled examples, which is used to compare a given testing input to the calibration set $\mathcal{D}_{\text{cal}}$. Since non-conformity score can be intuitively converted to a conformity measure (Vovk, Gammerman, and Shafer 2005), we use non-conformity measure for ease of technical exposition. A typical method based on split conformal prediction (see Algorithm 1) has a threshold parameter $\tau \to t$ to compute UQ in the form of prediction set for a given testing input x and deep model F_θ. A small set of calibration examples $\mathcal{D}_{\text{cal}}$ are used to select the threshold t for achieving the given coverage $1 - \alpha$ (say 90%) empirically on $\mathcal{D}_{\text{cal}}$. For example, in classification tasks, we select the t as $(1 - \alpha)$-quantile of $V(x, y^*)$ on the calibration set $\mathcal{D}_{\text{cal}}$ and the prediction set for a new testing input x is given by $\mathcal{C}(x)$=$\{y : V(x, y) \leq t\}$. CP provides formal guarantees that $\mathcal{C}(x)$ has coverage $1 - \alpha$ on a future testing input from the same distribution $\mathcal{D}_{\mathcal{Z}}$.

Algorithm 1: Split Conformal Prediction (CP)

1: **Input**: Significance level $\alpha \in (0, 1)$; Randomly split data into training set $\mathcal{D}_{\text{tr}}$ and calibration set $\mathcal{D}_{\text{cal}} = \{Z_1, \cdots, Z_n\}$.
2: If predictor F_θ is not given, train a prediction model F_θ on the training set $\mathcal{D}_{\text{tr}}$
3: Compute non-conformity score V_i for each example $Z_i \in \mathcal{D}_{\text{cal}}$
4: Compute $\hat{Q}^{\text{CP}}(\alpha, V_{1:n})$ as the $\lceil (1 - \alpha)(1 + |\mathcal{D}_{cal}|) \rceil$th smallest value in $\{V_i\}_{i \in \mathcal{D}_{\text{cal}}}$ as in (1).
5: $\hat{\mathcal{C}}(x_{n+1}) = \{y : V(x_{n+1}, y) \leq \hat{Q}^{\text{CP}}(\alpha, V_{1:n})\}$ is the prediction set for a testing input x_{n+1}

For classification, recent work has proposed conformity scores based on ordered probabilities (Romano, Sesia, and Candes 2020; Angelopoulos et al. 2021). The conformity score of adaptive prediction sets (APS) (Romano, Sesia, and Candes 2020) is defined as follows. For a given input x, we get the sorted probabilities for all classes using the deep model F_θ, $\pi(x, y^1) \geq \cdots \pi(x, y^C)$, and compute the score:

$$V^{\text{APS}}(x, k) = \pi(x, y^1) + \cdots + \pi(x, y^{k-1}) + U \cdot \pi(x, y^k)$$

where $U \in [0, 1]$ is a random variable to break ties. Suppose L is the index of the ground-truth class label y^* in the ordered list of probabilities $\pi(x, y)$. The conformity scoring function of regularized APS (RAPS) (Angelopoulos et al. 2021) is:

$$V^{\text{RAPS}}(x, k) = V_{APS}(x, k) + \lambda_R \cdot |L - k_{reg}|$$

where λ_R is the regularization parameter and k_{reg} is another parameter which is set based on the distribution of L values on validation data. Since NCP is a wrapper algorithm, we will employ both APS and RAPS to demonstrate the effectiveness of NCP in improving efficiency.

Problem definition. The high-level goal of this paper is to study provable methods to improve the standard CP framework to achieve high-efficiency (small prediction set) by meeting the coverage constraint $1-\alpha$. Specifically, we propose and study the neighborhood CP algorithm that assigns higher importance to calibration examples in the local neighborhood of a given testing example. We theoretically and empirically analyze when/why neighborhood CP algorithm results in improved efficiency.

3 Neighborhood Conformal Prediction

In this section, we provide the details of the Neighborhood CP (NCP) algorithm to improve the efficiency of CP. We start with some basic notation and terminology.

Let $V : \mathcal{Z} \to \mathbb{R}$ denote the non-conformity measure function. For each data sample Z_i in the calibration set $\mathcal{D}_{\text{cal}}$, we denote the corresponding non-conformity score as $V_i = V(Z_i)$. For a pre-specified level of coverage $1-\alpha$, we can determine the quantile on $\mathcal{D}_{\text{cal}}$ as shown below.

$$\hat{Q}^{\text{CP}}(\alpha, V_{1:n}) = \min\left\{t : \sum_{i=1}^{n} \frac{1}{n} \cdot \mathbf{1}[V_i \le t] \ge 1-\alpha\right\} \quad (1)$$

where t is the threshold parameter, $\mathbf{1}$ is the indicator function, and n is the number of calibration examples, i.e., n=$|\mathcal{D}_{\text{cal}}|$. We can measure the *efficiency* of conformal prediction using this quantile. Given a fixed α, a smaller quantile means a smaller threshold t to achieve the same significance level α, leading to more efficient uncertainty quantification (i.e., smaller prediction set). We provide a generic split conformal prediction algorithm in Algorithm 1 for completeness.

In many real-world applications, the conditional distribution of output y given input x can vary due to the heterogeneity in inputs. Thus, it is desirable to exploit this inherent heterogeneity during uncertainty quantification. The key idea behind the localized conformal prediction framework (Guan 2021) is to assign importance weights to conformal scores on calibration examples: higher importance is given to calibration samples in the local neighborhood of a given testing example X_{n+1}. In contrast, standard CP assigns uniform weights to calibration examples. We summarize the NCP procedure in Algorithm 2. Importantly, NCP is a wrapper approach in the sense that it can be used with any conformity scoring function. For example, if we use a better conformity score (say RAPS), we will get higher improvements in efficiency of UQ through NCP as demonstrated by our experiments.

Suppose $p_{i,j}$ denotes a weighting function that accounts for the similarity between any i-th and j-th calibration samples from $\mathcal{D}_{\text{cal}}$. To determine the NCP quantile given a mis-coverage level α, it suffices to find $\hat{\alpha}^{\text{NCP}}(\alpha)$ as below:

$$\hat{\alpha}^{\text{NCP}}(\alpha) = \max\{\tilde{\alpha} : \textstyle\sum_{i=1}^{n} \frac{1}{n}\mathbf{1}[V_i \le \hat{Q}^{\text{NCP}}(\tilde{\alpha}; V_{1:n}; p_{i,1:n})] \ge 1-\alpha\} \quad (2)$$

where the NCP quantile is determined as follows:

$$\hat{Q}^{\text{NCP}}(\tilde{\alpha}; V_{1:n}; p_{i,1:n}) = \min\{t : \sum_{j=1}^{n} p_{i,j}\mathbf{1}[V_j \le t] \ge 1-\tilde{\alpha}\}$$

Prior work on localized CP (Guan 2021) mainly focused on theoretical analysis of finite-sample coverage guarantees, synthetic experiments using raw inputs x for the regression task, and acknowledged that design of localizer and weighting function for practical purposes is out of their scope. Since the efficiency of UQ from LCP critically depends on the localizer and weighting function, our NCP algorithm specifies an effective localizer that leverages the learned input representation $\Phi(x)$ from the pre-trained deep model F_θ (e.g., the output of the layer before the softmax layer in CNNs for image classification) for our theoretical and empirical analysis. As we show in our theoretical analysis, input representation $\Phi(x)$ exhibiting good separation between inputs from different classes will result in small prediction sets.

Algorithm 2: Neighborhood Conformal Prediction (NCP)

1: **Input**: Significance level $\alpha \in (0,1)$. Randomly split data into training set $\mathcal{D}_{\text{tr}}$ and calibration set $\mathcal{D}_{\text{cal}} = \{Z_1, \cdots, Z_n\}$.
2: If predictor F_θ is not given, train a prediction model F_θ on the training set $\mathcal{D}_{\text{tr}}$.
3: Compute non-conformity score V_i for $Z \in \mathcal{D}_{\text{cal}}$
4: Compute importance weights $p_{i,j}$ for $X_i, X_j \in \mathcal{D}_{\text{cal}}$ needed to identify *neighborhood* according to (3 or 5)
5: Find α^{NCP} in (2) on $\mathcal{D}_{\text{cal}}$ and set $\hat{\mathcal{C}}(X_{n+1}) = \{y : V(X_{n+1}, y) \le \hat{Q}(\alpha^{\text{NCP}}, V_{1:n+1}, p_{n+1,1:n+1})\}$ as the prediction set for the new data sample X_{n+1}

Localizer and weighting function. We propose a ball-based localizer function for our NCP algorithm as shown below.

$$p_{i,j} = \frac{\mathbf{1}[\Phi(X_j) \in \mathcal{B}(\Phi(X_i))]}{\sum_{k \in \mathcal{D}_{\text{cal}}} \mathbf{1}[\Phi(X_k) \in \mathcal{B}(\Phi(X_i))]}, \quad (3)$$

where $\Phi(X)$ is the learned representations of X by the deep model F_θ, $\mathbf{1}$ is the indicator function, and $\mathcal{B}(x) \triangleq \{x' \in \mathcal{X} : \|x - x'\| \le B\}$ denotes a Euclidean ball with radius B centered at x. In the next section, we theoretically analyze the improved efficiency of NCP (Algorithm 2) over traditional CP algorithms under some mild conditions over the learned input representation by (pre-trained) deep classifier. In some scenarios, a Euclidean ball of fixed radius to define the neighborhood of a testing example may fail as we may not find any calibration examples in the neighborhood. Therefore, we propose the following weighting function $p_{i,j} \leftarrow p_{i,j}^{exp}$ over k nearest neighbors of a sample.

$$IW(x_i, x_j) = \exp\left(-\frac{dist(\Phi(x_i), \Phi(x_j))}{\lambda_L}\right) \quad (4)$$

$$p_{i,j}^{exp} = \frac{IW(x_i, x_j)}{\sum_{k \in \mathcal{D}_{cal}(KNN)} IW(x_i, x_k)} \quad (5)$$

where $dist(\Phi(x_i), \Phi(x_j))$ is a distance measure to capture the dissimilarity, which is the Euclidean distance in our case; λ_L is a tunable hyper-parameter (smaller value results in more localization); and non-zero importance weights are assigned to *only* the k nearest neighbors, i.e., $IW(x_i, x_j)$=0 if x_j is *not* one of the k nearest-neighbors of x_i. Our theoretical analysis is applicable to any localizer and weighting function defined in terms of $\Phi(.)$, but we employ the localizer based on Equation (5) for our experiments. Our ablation experiments (see Table 1 and Table 2) comparing NCP with a variant that includes all calibration examples in the neighborhood justify the appropriateness of selecting k nearest neighbors.

4 Theoretical Analysis

In this section, we theoretically analyze the efficiency of NCP when compared with standard CP. Recall that efficiency of a conformal prediction approach is related to the size of prediction set (small value means higher efficiency) for a pre-defined significance level α. Before providing the analysis, we first give our criterion for efficiency comparison. Since the quantile increases as α increases, we employ the following approach for comparing NCP and CP. We regard the quantile value as a function of the pre-defined significance level α. Specifically, for CP, there is a constant quantile for all data samples, i.e., $Q^{\text{CP}}(\alpha)$, while for NCP, the quantile is adaptive and depends on the data sample. Therefore, we use expected quantile value as the measure of efficiency of NCP, denoted by $\bar{Q}^{\text{NCP}}(\alpha)$. To directly compare the efficiency of both algorithms, we fix the same α and compare the two quantile functions. We first determine the quantile using CP for achieving the target significance level α. Next, we plug this CP quantile into NCP to show that NCP can always achieve higher coverage under some conditions on the distribution of learned input representation by deep classifier and the alignment with non-conformity measures on such distribution. The main reason for requiring synchronizing the quantile and to compare the coverage is that NCP may have adaptive quantile depending on the specific data sample, while CP has a fixed quantile for all samples. Hence, it is difficult to compare their quantiles given the same α.

We assume that the data samples from feature space $\mathcal{X}$ can be classified into C classes. We assume that data samples are drawn from distribution $\mathcal{D}_{\mathcal{X}}$ and the ground-truth labels are generated by a target function F^*[1] $: \mathcal{X} \to [C]$. For each $c \in [C]$, let $\mathcal{X}_c = \{x \in \mathcal{X} : F^*(x) = c\}$ denote the samples with ground truth label c and $P_{\text{class}}^{min} = \min_{c \in [C]} \mathbb{P}_X\{X \in \mathcal{X}_c\}$. Our analysis relies on the concept of neighborhood of data sample(s). $\mathcal{N}_B(x)$ be the neighborhood of x, i.e., $\mathcal{N}_{\mathcal{B}}(x) \triangleq \{x' : \mathcal{B}(\Phi(x)) \cap \mathcal{B}(\Phi(x')) \neq \emptyset\}$. We define a robust set of F^* by $\mathcal{R}_{\mathcal{B}}^* \triangleq \{x \in \mathcal{X} : F^*(x) = F^*(x'), \forall x' \in \mathcal{N}_{\mathcal{B}}(x)\}$, in which all elements share the same label with its neighbors. Below we introduce several important definitions.

[1]We used $F^*(x) = y^* = y$.

Definition 1. *(σ-concentration of quantiles on $\mathcal{R}_{\mathcal{B}}^* \cap \mathcal{X}_c$) Let $\sigma \geq 1$ be some constant. If for any quantile t of the non-conformity scores and any $X \in \mathcal{R}_{\mathcal{B}}^* \cap \mathcal{X}_c$ where $c \in [C]$, we have*

$$\mathbb{P}_{X'}\{V(X', F^*(X')) \leq t, X' \in \mathcal{N}_{\mathcal{B}}(X) | X \in \mathcal{R}_{\mathcal{B}}^* \cap \mathcal{X}_c\} \geq \sigma \cdot \mathbb{P}_{X'}\{V(X', F^*(X')) \leq t | X' \in \mathcal{X}_c\},$$

then the quantiles are σ-concentrated on $\mathcal{R}_{\mathcal{B}}^* \cap \mathcal{X}_c$.

Definition 1 states that samples satisfying any quantile of non-conformity scores V are distributed *significantly densely* on $\mathcal{R}_{\mathcal{B}}^* \cap \mathcal{X}_c$. Considering only those samples with less than t non-conformity scores and assuming they are distributed uniformly in the entire $\mathcal{X}_c$, the associated condition number $\sigma = 1$. If the distribution of such samples can be more concentrated on $\mathcal{R}_{\mathcal{B}}^* \cap \mathcal{X}_c$ rather than the non-robust set $\mathcal{X}_c \backslash \mathcal{R}_{\mathcal{B}}^*$, then we have a larger condition number $\sigma > 1$.

Definition 2. *($\mu_{\mathcal{B}}$-separation) If $\max_{c \in [C]} \mathbb{P}_X\{X \notin \mathcal{R}_{\mathcal{B}}^*, F^*(X) = c\} \leq \mu_{\mathcal{B}}$, then the underlying distribution $\mathcal{D}_{\mathcal{X}}$ is $\mu_{\mathcal{B}}$-separable.*

Definition 2 states that the region of robust set is lower bounded for each class, i.e., $\mathbb{P}_X\{X \in \mathcal{R}_{\mathcal{B}}^*, F^*(X) = c\} \geq 1 - \mu_{\mathcal{B}}$. A smaller condition number $\mu_{\mathcal{B}}$ means that more percentage of data samples are distributed in the robust set. The above definitions are built on population distribution rather than training set, as the expansion condition in (Wei et al. 2020).

Assumption 1. *Suppose α is the pre-defined significance level and: (i) any quantile t satisfies σ-concentration on $\mathcal{R}_{\mathcal{B}}^* \cap \mathcal{X}_c$ for $c \in [C]$; (ii) $\mathcal{D}_{\mathcal{X}}$ is $\mu_{\mathcal{B}}$ separable; (iii) the condition numbers in (i) and (ii) satisfy $(1 - \mu_{\mathcal{B}})\sigma \geq (1 - \alpha)/(1 - \alpha(2 - P_{\text{class}}^{min}))$; (iv) $\alpha \leq 1/2$.*

We make assumptions for any quantile t of non-conformity scores at the population level. Given the above definition of the neighborhood $\mathcal{N}_{\mathcal{B}}$, $\mu_{\mathcal{B}}$ implicitly describes the difficulty of the classification problem. The assumption on the concentration of quantiles requires that the non-conformity scores V are consistent with the underlying distribution of the input representation $\mathcal{D}_{\mathcal{X}}$. Using the concepts of robust set $\mathcal{R}_{\mathcal{B}}^*$ and its neighborhood region $\mathcal{N}_{\mathcal{B}}(\mathcal{R}_{\mathcal{B}}^*)$ (or its class-wise component $\mathcal{N}_{\mathcal{B}}(\mathcal{R}_{\mathcal{B}}^* \cap \mathcal{X}_c)$ as in Definition 1), we can partition the entire sample space $\mathcal{X}$ into two disjoint parts: *core part* (robust set), and the *remaining part* (outside the robust set). Good input representation $\mathcal{X}$ will achieve good separation of data clusters for different classes, i.e., the above condition number $(1 - \mu_{\mathcal{B}})\sigma$ can be larger than 1.

Theorem 1. *Suppose Assumption 1 holds. Let $Q^{CP}(\alpha) \triangleq \min\{t : \mathbb{P}_X\{V(X, F^*(X)) \leq t\} \geq 1 - \alpha\}$, $\alpha^{NCP}(\alpha) \triangleq \max\{\tilde{\alpha} : \mathbb{P}_X\{X \leq Q^{NCP}(\tilde{\alpha}; X)\} \geq 1 - \alpha\}$, where $Q^{NCP}(\tilde{\alpha}; X) \triangleq \min\{t : \mathbb{P}_{X'}\{V(X'; F^*(X)) \leq t, X' \in \mathcal{N}_{\mathcal{B}}(X)\} \geq 1 - \tilde{\alpha}\}$ in population. Then to achieve the same α, NCP gives smaller expected quantile and can be more efficient than CP: $\mathbb{E}_X[Q^{NCP}(\alpha; X)] \leq Q^{CP}(\alpha)$.*

Proof. (Sketch) We present the complete proof in Appendix due to space constraints. Our theoretical analysis proceeds with two key milestones. In the first step, we establish a connection from NCP to an artificial CP algorithm referred

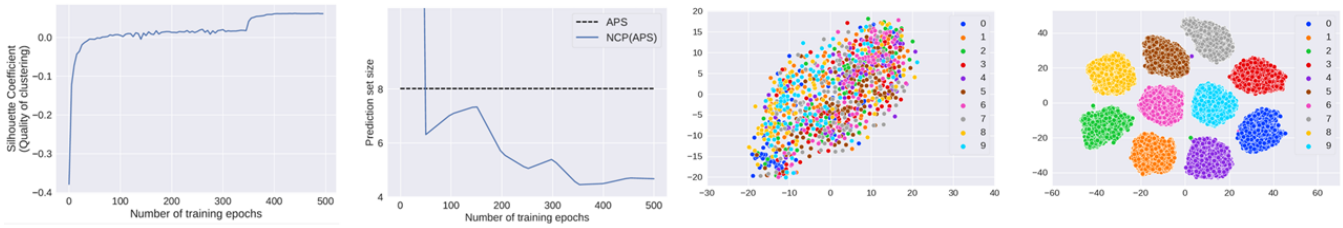

Figure 1: Ablation results to justify NCP algorithm and its theoretical analysis. (**Left**) As a function of the number of training epochs for ResNet50 trained on CIFAR100, we show (a) the clustering property of learned input representation and (b) prediction set size from NCP. The conformity score is kept constant for both CP and NCP. The prediction set size from NCP reduces over CP as the clustering property of the learned representation improves. (**Right**) the t-SNE visualization of CIFAR10 data categorized by their ground-truth class labels using (c) raw image pixels and (d) learned input representation from ResNet50 model after 500 training epochs. The learned representation exhibits significantly a better clustering property.

to as *class-wise NCP* that determines the threshold for each class independently. The gap between NCP and class-wise NCP can be filled by using the assumed σ-concentration and $\mu_{\mathcal{B}}$-separation, since both conditions characterize the properties of input representation along different dimensions. Our analysis shows a clear improvement of efficiency by NCP, i.e., smaller expected quantile over the data distribution when compared to class-wise NCP. In the second step, we build a link between class-wise NCP with traditional CP. Even though they can both achieve marginal coverage, class-wise NCP finds the quantiles for individual classes *adaptively* (i.e., quantiles may vary for different classes). However, traditional CP can only determine a *global* quantile across the entire data distribution. Finally, we combine these two milestones to connect NCP to traditional CP to show improved efficiency.

Remark 1. The above result treats the quantiles determined by NCP and CP as a function of α. When achieving the same α, the expected quantile derived by NCP is smaller than the one determined by CP. The key idea behind this result is to make use of the alignment between the quantiles and distribution, which critically depends on the learned input representation of $\mathcal{X}$ by deep classifiers. Learned input representations with a good clustering property will reduce the prediction set size. The sample distribution is characterized by the robust set $\mathcal{R}^*_{\mathcal{B}}$ and the remaining region. Additionally, if the non-conformity scores are sufficiently consistent with data distribution in any quantiles, then a well-designed localizer can reduce the required quantile to achieve α, the target significance level.

5 Experiments and Results

We present experimental evaluation of NCP over prior CP methods for classification and other baselines to demonstrate the effectiveness of NCP in reducing prediction set size.

5.1 Experimental Setup

Classification datasets. We consider CIFAR10 (Krizhevsky, Hinton et al. 2009), CIFAR100 (Krizhevsky, Hinton et al. 2009), and ImageNet (Deng et al. 2009) datasets using the standard training and testing set split. We employ the same methodology as (Angelopoulos et al. 2021) to create calibration data and validation data for tuning hyper-parameters.

Table 1: The average percentage of calibration examples used by NCP as nearest neighbors to produce prediction sets.

Model	ImageNet	CIFAR100	CIFAR10
ResNet18	24.8	12.3	09.0
ResNet50	12.8	08.3	15.0
ResNet101	33.2	12.0	13.0

Deep models for classification. We train three commonly used ResNet architectures, namely, ResNet18, ResNet50, and ResNet101 (He et al. 2016) on CIFAR10 and CIFAR100. For ImageNet, we employ nine pre-trained deep models similar to (Angelopoulos et al. 2021), namely, ResNeXt101, ResNet152, ResNet101, ResNet50, ResNet18, DenseNet161, VGG16, Inception, and ShuffleNet from the TorchVision repository (Paszke et al. 2019). We calibrate classifiers via Platt scaling (Guo et al. 2017) on the calibration data before applying CP methods.

Methods and baselines. We consider the conformity score of APS (Romano, Sesia, and Candès 2020) and RAPS (Angelopoulos et al. 2021) as strong CP baselines. Since NCP is a wrapper algorithm that can work with any conformity score, our experiments will demonstrate the efficacy of NCP to improve over both APS and RAPS, namely, NCP(APS) and NCP(RAPS). We employ the publicly available implementation[2] of APS and RAPS, and build on it to implement our NCP algorithm. We also compare with a naive baseline (Angelopoulos et al. 2021) that produces prediction set by including classes from highest to lowest probability until their cumulative sum just exceeds the threshold $1-\alpha$.

Evaluation methodology. We select the hyper-parameters for RAPS and NCP from the following values noting that Bayesian optimization (Shahriari et al. 2016; Snoek, Larochelle, and Adams 2012) can be used for im-

[2]https://github.com/aangelopoulos/conformal_classification/

proved results: $\lambda_L = \{10, 50, 100, 500, 1000, 5000\}$ and $\lambda_R = \{0.001, 0.005, 0.01, 0.05, 0.15, 0.2, 0.3, 0.4, 0.5, 1.0\}$. We create calibration and validation data following the methodology in (Angelopoulos et al. 2021) and use validation data for tuning the hyper-parameters. We present all our experimental results for desired coverage as 90% (unless specified otherwise). We compute two metrics computed on the testing set: *Coverage* (fraction of testing examples for which prediction set contains the ground-truth output) and *Efficiency* (average length of cardinality of prediction set for classification, small values mean high efficiency). We report the average metrics over ten different runs for CIFAR100 and CIFAR10, and five different runs for ImageNet.

5.2 Results and Discussion

Empirical support for the theory of `NCP`. Figure 1 (a) and (b) shows that as the clustering property of the learned input representation improves, prediction set size from `NCP(APS)` reduces and improves over APS. This empirical result clearly demonstrates our theoretical result comparing `NCP` and CP. Figure 1 (c) and (d) show the t-SNE visualization of raw image pixels and learned input representation from ResNet50 on CIFAR10 data. This result demonstrates that learned representation exhibits significantly better clustering property over raw data and justifies its use for the `NCP`'s localizer.

Ablation results for `NCP`. Table 1 shows the fraction of calibration examples used by `NCP(APS)` on an average to compute prediction sets for three ResNet models noting that we find similar patterns for other deep models. These results demonstrate that a small fraction of calibration examples are used by `NCP(APS)` as neighborhood. One interesting ablation question is how does the `NCP` variant using all calibration examples compare to `NCP`? To answer this question, Table 2 shows the mean prediction set size for `NCP(APS)-All` and `NCP(APS)`. The results clearly demonstrate that `NCP(APS)` produces significantly smaller predicted set sizes justifying the selection of k nearest neighbors in `NCP`.

`NCP` vs. Baseline methods. Table 3 and Table 4 show respectively the results for efficiency (average prediction set size) and marginal coverage for ImageNet, CIFAR100, and CIFAR10 datasets respectively. We discuss all these results along different dimensions and make the following observations. **1)** The prediction set size is significantly reduced by `NCP(APS)` for both ImageNet and CIFAR100, compared to the APS and Naive algorithms. The prototypical results shown in Figure 1 demonstrate that the prediction set size from `NCP` is directly proportional to the clustering property of the learned input representation. **2)** `NCP(RAPS)` reduces the prediction set size over APS, RAPS, and Naive baselines. These results demonstrate that `NCP` being a wrapper algorithm can make use of a better conformity scoring function (i.e., RAPS) to further reduce the prediction set size. **3)** The reduction in size of prediction set by `NCP(APS)` and `NCP(RAPS)` is achieved by further tightening the gap between actual coverage and the desired coverage $1 - \alpha$. **4)** Naive baseline achieves the desired coverage, but the size of its prediction set is significantly larger than all CP methods.

6 Related Work

Conformal prediction (CP) is a general framework for uncertainty quantification that provides (marginal) coverage guarantees without any assumptions on the underlying data distribution (Vovk, Gammerman, and Saunders 1999; Vovk, Gammerman, and Shafer 2005; Shafer and Vovk 2008). Various instantiations of the CP framework are studied to produce prediction intervals for regression (Papadopoulos 2008; Vovk 2012, 2015; Lei and Wasserman 2014; Vovk et al. 2018; Lei et al. 2018; Romano, Patterson, and Candes 2019; Izbicki, Shimizu, and Stern 2019; Guan 2019; Gupta, Kuchibhotla, and Ramdas 2022; Kivaranovic, Johnson, and Leeb 2020; Barber et al. 2021; Foygel Barber et al. 2021), and prediction sets for multi-class classification (Lei, Robins, and Wasserman 2013; Sadinle, Lei, and Wasserman 2019; Romano, Sesia, and Candès 2020; Angelopoulos et al. 2021) and structured prediction (Bates et al. 2021).

This paper focuses on split conformal prediction (Papadopoulos 2008; Lei et al. 2018), which uses a set of calibration examples to provide uncertainty quantification for any given predictor including deep models. Our problem setup considers CP methods to reduce prediction set size under a given marginal coverage constraint. There is relatively less work on CP methods for classification when compared to regression. The adaptive prediction sets (APS) method (Romano, Sesia, and Candès 2020) and its regularized version (referred to as RAPS) (Angelopoulos et al. 2021) design conformity scoring functions to return smaller prediction sets. However, neither APS nor RAPS were analyzed theoretically to characterize the reasons for its effectiveness. Our `NCP` algorithm is a wrapper approach that is complementary to both APS and RAPS as we demonstrated in experiments.

Localized CP (Guan 2021) was proposed to improve CP by putting more emphasis on the local neighborhood of testing examples and was only evaluated on synthetic regression tasks. Recent work (Lin, Trivedi, and Sun 2021) applied LCP for real-world regression tasks with good experimental results. *However, there is no existing theoretical analysis to characterize the precise conditions for reduced prediction intervals or prediction sets.* In fact, this analysis is not possible without defining a concrete localizer function. The proposed neighborhood CP algorithm specifies an effective localizer function by leveraging the input representations learned by deep models, develops theory to characterize when and why neighborhood CP produces smaller prediction sets for classification problems, and performs empirical evaluation on real-world classification datasets using diverse deep models to demonstrate the effectiveness of this simple approach.

There are also other approaches which are not based on conformal prediction to produce prediction sets with small sizes for regression (Pearce et al. 2018; Chen et al. 2021) and classification (Park et al. 2019) tasks. Since the focus of this paper is on improving CP based uncertainty quantification, these methods are out of scope for our study.

7 Summary

This paper studied a novel neighborhood conformal prediction (`NCP`) algorithm to improve uncertainty quantification

Table 2: Ablation results comparing `NCP` and `NCP` variant using all calibration examples as neighborhood `(NCP-All)`. `NCP` reduces the mean prediction set size by using a smaller neighborhood. Both `NCP` and `NCP-All` achieve $1-\alpha$ coverage.

Dataset	ImageNet		CIFAR100		CIFAR10	
$1-\alpha$	0.90		0.90		0.96	
	NCP-All	NCP	NCP-All	NCP	NCP-All	NCP
ResNet18	11.82	**11.08**	4.87	**3.61**	1.22	**1.07**
ResNet50	09.24	**06.44**	7.45	**4.26**	1.16	**1.05**
ResNet101	07.90	**05.60**	4.17	**2.80**	1.18	**1.04**

Table 3: **Mean prediction set size** on ImageNet, CIFAR100, and CIFAR10. We present the mean and standard deviation over five different runs for ImageNet and ten different runs for CIFAR100 and CIFAR10 respectively.

Model	Naive	APS	RAPS	NCP(APS)	NCP(RAPS)
ImageNet($1-\alpha=0.900$)					
ResNet152	10.55(0.548)	11.34(0.843)	2.53(0.054)	5.24(0.312)	**2.12(0.007)**
ResNet101	10.85(0.588)	11.36(0.615)	2.63(0.08)	5.60(0.278)	**2.25(0.098)**
ResNet50	12.00(0.530)	13.24(0.892)	2.94(0.089)	6.44(0.381)	**2.54(0.091)**
ResNet18	16.13(0.642)	17.05(1.100)	5.00(0.220)	11.08(0.598)	**4.71(0.251)**
ResNeXt101	18.57(1.05)	20.57(1.10)	2.36(0.069)	8.24(0.388)	**2.01(0.060)**
DenseNet161	11.70(0.730)	12.86(1.21)	2.67(0.074)	5.89(0.447)	**2.29(0.103)**
VGG16	13.91(0.867)	14.10(0.486)	3.97(0.098)	8.56(0.382)	**3.62(0.106)**
Inception	77.51(3.78)	90.28(4.16)	5.96(0.373)	66.29(2.83)	**5.67(0.178)**
ShuffleNet	30.33(1.64)	34.31(2.10)	5.53(0.070)	22.36(1.37)	**5.49(0.131)**
CIFAR100 ($1-\alpha=0.900$)					
ResNet18	5.07(0.327)	5.57(0.224)	3.17(0.076)	3.61(0.164)	**2.77(0.100)**
ResNet50	8.21(0.746)	8.02(0.405)	3.37(0.189)	4.26(0.233)	**2.96(0.213)**
ResNet101	4.44(0.348)	4.64(0.184)	2.60(0.066)	2.80(0.188)	**2.19(0.073)**
CIFAR10($1-\alpha=0.960$)					
ResNet18	1.20(0.016)	1.25(0.021)	1.24(0.014)	1.07(0.013)	**1.06(0.012)**
ResNet50	1.15(0.013)	1.19(0.021)	1.17(0.011)	1.05(0.010)	**1.04(0.005)**
ResNet101	1.17(0.017)	1.20(0.018)	1.19(0.017)	1.04(0.008)	**1.03(0.007)**

Table 4: **Marginal coverage** on ImageNet, CIFAR100, and CIFAR10. We present the mean and standard deviation over five different runs for ImageNet and ten different runs for CIFAR100 and CIFAR10 respectively.

Model	Naive	APS	RAPS	NCP(APS)	NCP(RAPS)
ImageNet($1-\alpha=0.900$)					
ResNet152	0.937(0.002)	0.940(0.003)	0.922(0.003)	0.905(0.005)	0.904(0.007)
ResNet101	0.936(0.002)	0.938(0.002)	0.919(0.003)	0.903(0.004)	0.905(0.008)
ResNet50	0.934(0.001)	0.938(0.003)	0.918(0.004)	0.902(0.005)	0.904(0.007)
ResNet18	0.925(0.003)	0.928(0.004)	0.910(0.003)	0.901(0.006)	0.903(0.006)
ResNeXt101	0.935(0.001)	0.939(0.002)	0.920(0.003)	0.905(0.003)	0.903(0.006)
DenseNet161	0.934(0.001)	0.937(0.003)	0.919(0.003)	0.902(0.007)	0.904(0.007)
VGG16	0.929(0.003)	0.930(0.003)	0.912(0.003)	0.906(0.004)	0.904(0.004)
Inception	0.919(0.002)	0.927(0.004)	0.909(0.004)	0.910(0.003)	0.905(0.003)
ShuffleNet	0.927(0.003)	0.932(0.004)	0.907(0.002)	0.911(0.005)	0.907(0.003)
CIFAR100 ($1-\alpha=0.900$)					
ResNet18	0.931(0.005)	0.938(0.004)	0.925(0.003)	0.908(0.007)	0.909(0.006)
ResNet50	0.947(0.006)	0.944(0.005)	0.919(0.004)	0.906(0.008)	0.906(0.011)
ResNet101	0.940(0.005)	0.944(0.004)	0.926(0.004)	0.907(0.008)	0.907(0.006)
CIFAR10($1-\alpha=0.960$)					
ResNet18	0.979(0.002)	0.982(0.002)	0.982(0.002)	0.962(0.005)	0.961(0.004)
ResNet50	0.981(0.003)	0.984(0.003)	0.984(0.003)	0.965(0.003)	0.963(0.003)
ResNet101	0.983(0.002)	0.986(0.002)	0.985(0.003)	0.966(0.002)	0.964(0.002)

of pre-trained deep classifiers. For a given testing input, NCP identifies k nearest neighbors and assigns importance weights proportional to their distance defined using the learned representation of deep classifier. The theoretical analysis characterized why NCP reduces the prediction set size over standard CP framework. Our experimental results corroborated the developed theory and demonstrated significant reduction in prediction set size over prior CP methods on diverse classification benchmarks and deep models.

Acknowledgments

This research is supported in part by Proofpoint Inc. and the AgAID AI Institute for Agriculture Decision Support, supported by the National Science Foundation and United States Department of Agriculture - National Institute of Food and Agriculture award #2021-67021-35344.

References

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(ICLR)*.

Barber, R. F.; Candes, E. J.; Ramdas, A.; and Tibshirani, R. J. 2021. Predictive inference with the jackknife+. *The Annals of Statistics*.

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

Chen, H.; Huang, Z.; Lam, H.; Qian, H.; and Zhang, H. 2021. Learning prediction intervals for regression: Generalization and calibration. In *International Conference on Artificial Intelligence and Statistics(AISTATS)*. PMLR.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE conference on Computer Vision and Pattern Recognition(CVPR)*.

Foygel Barber, R.; Candes, E. J.; Ramdas, A.; and Tibshirani, R. J. 2021. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*.

Guan, L. 2019. Conformal prediction with localization. *arXiv preprint arXiv:1908.08558*.

Guan, L. 2021. Localized Conformal Prediction: A Generalized Inference Framework for Conformal Prediction. *arXiv preprint arXiv:2106.08460*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning(ICML)*. PMLR.

Gupta, C.; Kuchibhotla, A. K.; and Ramdas, A. 2022. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE conference on Computer Vision and Pattern Recognition(CVPR)*.

Izbicki, R.; Shimizu, G. T.; and Stern, R. B. 2019. Flexible distribution-free conditional predictive bands using density estimators. *arXiv preprint arXiv:1910.05575*.

Kivaranovic, D.; Johnson, K. D.; and Leeb, H. 2020. Adaptive, distribution-free prediction intervals for deep networks. In *International Conference on Artificial Intelligence and Statistics(AISTATS)*. PMLR.

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

Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R. J.; and Wasserman, L. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*.

Lei, J.; Robins, J.; and Wasserman, L. 2013. Distribution-free prediction sets. *Journal of the American Statistical Association*.

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

Lin, Z.; Trivedi, S.; and Sun, J. 2021. Locally Valid and Discriminative Prediction Intervals for Deep Learning Models. *Advances in Neural Information Processing Systems(NeurIPS)*, 34.

Papadopoulos, H. 2008. Inductive Conformal Prediction: Theory and Application to Neural Networks. In Fritzsche, P., ed., *Tools in Artificial Intelligence*, chapter 18. Rijeka: IntechOpen.

Park, S.; Bastani, O.; Matni, N.; and Lee, I. 2019. PAC confidence sets for deep neural networks via calibrated prediction. *arXiv preprint arXiv:2001.00106*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems(NeurIPS)*, 32: 8026–8037.

Pearce, T.; Brintrup, A.; Zaki, M.; and Neely, A. 2018. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International Conference on Machine Learning(ICML)*. PMLR.

Romano, Y.; Patterson, E.; and Candes, E. 2019. Conformalized quantile regression. *Advances in Neural Information Processing Systems(NeurIPS)*.

Romano, Y.; Sesia, M.; and Candes, E. 2020. Classification with Valid and Adaptive Coverage. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems(NeurIPS)*, volume 33, 3581–3591. Curran Associates, Inc.

Romano, Y.; Sesia, M.; and Candès, E. J. 2020. Classification with valid and adaptive coverage. *arXiv preprint arXiv:2006.02544*.

Sadinle, M.; Lei, J.; and Wasserman, L. 2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*.

Shafer, G.; and Vovk, V. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research*.

Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; and De Freitas, N. 2016. Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems(NeurIPS)*.

Vovk, V. 2012. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning(ACML)*. PMLR.

Vovk, V. 2015. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*.

Vovk, V.; Gammerman, A.; and Saunders, C. 1999. Machine-learning applications of algorithmic randomness.

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

Vovk, V.; Nouretdinov, I.; Manokhin, V.; and Gammerman, A. 2018. Cross-conformal predictive distributions. In *Conformal and Probabilistic Prediction and Applications*. PMLR.

Wei, C.; Shen, K.; Chen, Y.; and Ma, T. 2020. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*.