PREDICTIVE INFERENCE WITH FEATURE CONFORMAL PREDICTION

Jiaye Teng[1,3,4,*], **Chuan Wen**[1,3,4,*], **Dinghuai Zhang**[2,*],
Yoshua Bengio[2], **Yang Gao**[1,3,4], **Yang Yuan**[1,3,4,†]
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University
[2]Mila - Quebec AI Institute
[3]Shanghai Artificial Intelligence Laboratory
[4]Shanghai Qi Zhi Institute
{tjy20,cwen20}@mails.tsinghua.edu.cn, dinghuai.zhang@mila.quebec

ABSTRACT

Conformal prediction is a distribution-free technique for establishing valid prediction intervals. Although conventionally people conduct conformal prediction in the output space, this is not the only possibility. In this paper, we propose feature conformal prediction, which extends the scope of conformal prediction to semantic feature spaces by leveraging the inductive bias of deep representation learning. From a theoretical perspective, we demonstrate that feature conformal prediction provably outperforms regular conformal prediction under mild assumptions. Our approach could be combined with not only vanilla conformal prediction, but also other adaptive conformal prediction methods. Apart from experiments on existing predictive inference benchmarks, we also demonstrate the state-of-the-art performance of the proposed methods on *large-scale* tasks such as ImageNet classification and Cityscapes image segmentation. The code is available in https://github.com/AlvinWen428/FeatureCP.

1 INTRODUCTION

Although machine learning models work well in numerous fields (Silver et al., 2017; Devlin et al., 2019; Brown et al., 2020), they usually suffer from over-confidence issues, yielding unsatisfactory uncertainty estimates (Guo et al., 2017a; Chen et al., 2021; Gawlikowski et al., 2021). To tackle the uncertainty issues, people have developed a multitude of uncertainty quantification techniques, including calibration (Guo et al., 2017b; Minderer et al., 2021), Bayesian neural networks (Smith, 2014; Blundell et al., 2015), and many others (Sullivan, 2015).

Among different uncertainty quantification techniques, *conformal prediction* (CP) stands out due to its simplicity and low computational cost properties (Vovk et al., 2005; Shafer & Vovk, 2008; Angelopoulos & Bates, 2021). Intuitively, conformal prediction first splits the dataset into a training fold and a calibration fold, then trains a machine learning model on the training fold, and finally constructs the confidence band via a non-conformity score on the calibration fold. Notably, the confidence band obtained by conformal prediction is *guaranteed* due to the exchangeability assumption in the data. With such a guarantee, conformal prediction has been shown to perform promisingly on numerous realistic applications (Lei & Candès, 2021b; Angelopoulos et al., 2022).

Despite its remarkable effectiveness, vanilla conformal prediction (vanilla CP) is only deployed in the output space, which is not the only possibility. As an alternative, feature space in deep learning stands out due to its powerful inductive bias of deep representation. Take the image segmentation problem as an example. In such problems, we anticipate a predictive model to be certain in the informative regions (*e.g.*, have clear objects), while uncertain elsewhere. Since different images would possess different object boundary regions, it is inappropriate to return the same uncertainty for different positions, as standard conformal prediction does. Nonetheless, if we instead employ conformal

[*]Equal Contribution.
[†]Correspond to yuanyang@mail.tsinghua.edu.

Figure 1: Illustration of vanilla CP (left) vs Feature CP (right). Feature CP operates in the semantic feature space, as opposed to the commonly adopted output space. These methods are described in further detail in Sections 3 and 4.

prediction on the more meaningful feature space, albeit all images have the same uncertainty on this intermediate space, the pixels would exhibit effectively different uncertainty in the output space after a non-trivial non-linear transformation (see Figure 3).

In this work, we thus propose the Feature Conformal Prediction (Feature CP) framework, which deploys conformal prediction in the feature space rather than the output space (see Figure 1). However, there are still two issues unsolved for performing Feature CP: (a) commonly used non-conformity scores require a ground truth term, but here the ground truth in feature space is not given; and (b) transferring the con dence band in the feature space to the output space is non-trivial. To solve problem (a), we propose a new non-conformity score based on the notation surrogate feature, which replaces the ground truth term in previous non-conformity scores. As for (b), we propose two methods: Band Estimation, which calculates the upper bound of the con dence band, together with Band Detection, to determine whether a response locates in the con dence band. More interestingly, feature-level techniques are pretty general and can be deployed into other distribution-free inference algorithms, e.g., conformalized quantile regression (CQR). This shows the great potential application impact of the proposed Feature CP methodology (see the discussion in Appendix B.4).

From a theoretical perspective, we demonstrate that Feature CP is provably more ef cient, in the sense that it yields shorter con dence bands than vanilla CP, given that the feature space meets cubic conditions. Here the cubic conditions sketch the properties of feature space from three perspectives, including length preserving, expansion, and quantile stability (see Theorem 6). At a colloquial level, the cubic conditions assume the feature space has a smaller distance between individual non-conformity scores and their quantiles, which reduces the cost of the quantile operation. We empirically validate that the feature space in deep learning satis es the cubic conditions, thus resulting in a better con dence band with a shorter length (See Figure 2) according to our theoretical analysis.

Our contributions can be summarized as follows:

- We propose Feature CP, together with a corresponding non-conformity score and an uncertainty band estimation method. The proposed method no longer treats the trained model as a black box but exploits the semantic feature space information. What's more, our approach could be directly deployed with any pretrained model as a plug-in component, without the need of re-training under specially designed learning criteria.
- Theoretical evidence guarantees that Feature CP is both (a) ef cient, where it yields shorter con dence bands, and (b) effective, where the empirical coverage provably exceeds the given con dence level, under reasonable assumptions.
- We conduct extensive experiments under both synthetic and realistic settings (e.g., pixel-level image segmentation) to corroborate the effectiveness of the proposed algorithm. Besides, we demonstrate the universal applicability of our method by deploying feature-level operations to improve other adaptive conformal prediction methods such as CQR.

2 RELATED WORK

Conformal prediction is a statistical framework dealing with uncertainty quantification problems (Vovk et al., 2005; Shafer & Vovk, 2008; Nouretdinov et al., 2011; Barber et al., 2020; Angelopoulos & Bates, 2021). The research on conformal prediction can be roughly split into the following branches. The first line of work focuses on relaxing the assumptions about data distribution in conformal prediction, e.g., exchangeability (Tibshirani et al., 2019; Hu & Lei, 2020; Podkopaev & Ramdas, 2021; Barber et al., 2022). The second line aims at improving the efficiency of conformal prediction (Romano et al., 2020b; Sesia & Candès, 2020; Izbicki et al., 2020a; Yang & Kuchibhotla, 2021; Stutz et al., 2021). The third line tries to generalize conformal prediction to different settings, e.g., quantile regression (Romano et al., 2019), K-Nearest Neighbors (Papadopoulos et al., 2011), density estimator (Izbicki et al., 2020b), survival analysis (Teng et al., 2021), causal (Lei et al., 2021), or conditional histogram regression (Sesia & Romano, 2021). There are also works combining conformal prediction with other machine learning topics, such as functional data (Lei et al., 2013), treatment effects (Lei & Candès, 2021a), time series analysis (Xu & Xie, 2021), online learning (Gibbs & Candès, 2021), adversarial robustness (Gendler et al., 2022), and many others.

Besides conformal prediction, there are many other uncertainty quantification techniques, including calibration (Guo et al., 2017a; Kuleshov et al., 2018; Nixon et al., 2019) and Bayesian-based techniques (Blundell et al., 2015; Hernández-Lobato & Adams, 2015; Li & Gal, 2017). Different from the above techniques, conformal prediction is appealing due to its simplicity, computationally free, and model-free properties.

Image segmentation is a traditional task in computer vision, which focuses on partitioning images into different semantic segments (Haralick & Shapiro, 1985; Senthilkumaran & Rajesh, 2009; Minaee et al., 2020). A line of researches applies conformal prediction with some threshold output for all pixels (Angelopoulos & Bates, 2021; Bates et al., 2021), or focus on the risk control tasks (Angelopoulos et al., 2021a). Different from previous approaches, our method first achieves meaningful pixel-level conformal prediction results to the best of our knowledge.

3 PRELIMINARIES

Predictive inference. Let $(X, Y) \sim P$ denotes a random data pair, e.g., an image and its segmentation map. Given a significance level α, we aim to construct a confidence band $C_{1-\alpha}(X)$, such that

$$P(Y \in C_{1-\alpha}(X)) \geq 1 - \alpha. \tag{1}$$

There is a tradeoff between efficiency and effectiveness, since one can always set $C_{1-\alpha}(X)$ to be infinitely large to satisfy Equation (1). In practice, we wish the measure of the confidence band (e.g., its length) can be as small as possible, given that the coverage in Equation (1) holds.

Dataset. Let $\mathcal{D} = \{(X_i, Y_i)\}_{i \in \mathcal{I}}$ denotes the dataset, where $\mathcal{I}$ denotes the set of data index and $(X_i, Y_i) \sim P$ denotes a sample pair following the distribution P. Typically, conformal prediction requires that data in $\mathcal{D}$ satisfies exchangeability (see below) rather than the stronger i.i.d. (independent and identically distributed) condition. We use $|\cdot|$ to represent the cardinality of a set. Conformal prediction needs to first randomly split the dataset into a training fold $\mathcal{D}_{tr} = \{(X_i, Y_i)\}_{i \in \mathcal{I}_{tr}}$ and a calibration fold $\mathcal{D}_{ca} = \{(X_i, Y_i)\}_{i \in \mathcal{I}_{ca}}$, where $\mathcal{I}_{tr} \cup \mathcal{I}_{ca} = \mathcal{I}$ and $\mathcal{I}_{tr} \cap \mathcal{I}_{ca} = \emptyset$. We denote the test point as (X^0, Y^0), which is also sampled from the distribution P.

Training process. During the training process, we train a machine learning model denoted by $\hat{\mu}(\cdot)$ (e.g., neural network) with the training fold $\mathcal{D}_{tr}$. For the ease of the following discussion, we rewrite the model as $\hat{\mu} = \hat{g} \circ \hat{f}$, where $\hat{f}$ denotes the feature function (i.e., first several layers in neural networks) and $\hat{g}$ denotes the prediction head (i.e., last several layers in neural networks).

Calibration process. Different from usual machine learning methods, conformal prediction has an additional calibration process. Specifically, we calculate the non-conformity score $V_i = s(X_i, Y_i, \hat{\mu})$ based on the calibration fold $\mathcal{D}_{ca}$, where $s(\cdot, \cdot, \cdot)$ is a function informally measuring how the model fits the ground truth. The simplest form of non-conformity score is $s(X_i, Y_i, \hat{\mu}) = \|Y_i - \hat{\mu}(X_i)\|$. One could adjust the form of the non-conformity score according to different contexts (e.g., Romano et al. (2019); Teng et al. (2021)). Based on the selected non-conformity score, a matching confidence band could be subsequently created.

Algorithm 1 Conformal Prediction

Require: Desired con dence level , dataset $D = \{(X_i, Y_i)\}_{i \in I}$, test point X^0, non-conformity score function $s(\cdot)$

1: Randomly split the dataset D into a training fold D_{tr}, $(X_i, Y_i)_{i \in I_{tr}}$ and a calibration fold D_{ca}, $(X_i, Y_i)_{i \in I_{ca}}$;
2: Train a base machine learning model $\mathcal{A}(\cdot, D_{tr})$ with D_{tr} to estimate the response Y;
3: For each $i \in I_{ca}$, calculate its non-conformity score $V_i = s(X_i, Y_i, \hat{\mu})$;
4: Calculate the $(1 -)$-th quantile Q_1 of the distribution $\frac{1}{|I_{ca}|+1} \sum_{i \in I_{ca}} V_i + 1$.

Ensure: $C_1(X^0) = \{Y : s(X^0, Y, \hat{\mu}) \leq Q_1\}$.

We present vanilla CP[1] in Algorithm 1. Moreover, we demonstrate its theoretical guarantee in Proposition 2, based on the following notation of exchangeability in Assumption 1.

Assumption 1 (exchangeability.) Assume that the calibration data $(X_i, Y_i), i \in I_{ca}$ and the test point (X^0, Y^0) are exchangeable. Formally, de ne $Z_i = (X_i, Y_i), i = 1, \ldots, |I_{ca}| + 1|$, as the above data pair, then Z_i are exchangeable if arbitrary permutation leads to the same distribution, i.e.,

$$(Z_1, \ldots, Z_{|I_{ca}|+1}) \overset{d}{=} (Z_{(1)}, \ldots, Z_{(|I_{ca}|+1)}), \tag{2}$$

with arbitrary permutation over $\{1, \ldots, |I_{ca} + 1|\}$.

Note that Assumption 1 is weaker than the i.i.d. assumption. Therefore, it is reasonable to assume the exchangeability condition to hold in practice. Based on the exchangeability assumption, one can show the following theorem, indicating that conformal prediction indeed returns a valid con dence band, which satis es Equation (1).

Theorem 2 (theoretical guarantee for conformal prediction (Law, 2006; Lei et al., 2018; Tibshirani et al., 2019).) Under Assumption 1, the con dence band $C_1(X^0)$ returned by Algorithm 1 satis es

$$P(Y^0 \in C_1(X^0)) \geq 1 - .$$

4 METHODOLOGY

In this section, we broaden the concept of conformal prediction using feature-level operations. This extends the scope of conformal prediction and makes it more exible. We analyze the algorithm components and details in Section 4.1 and Section 4.2. The algorithm is nally summarized in Section 4.3. We remark that although in this work we discuss Feature CP under regression regimes for simplicity's sake, one can easily extend the idea to classi cation problems.

4.1 NON-CONFORMITY SCORE

Conformal prediction necessitates a non-conformity score to measure the conformity between prediction and ground truth. Traditional conformal prediction usually uses norm-based non-conformity score due to its simplicity, i.e., $s(X, Y,) = \|Y - (X)\|$, where Y is the provided ground truth target label. Nonetheless, we have no access to the given target features if we want to conduct conformal prediction at the feature level. To this end, we introduce the surrogate feature (see De nition 3), which could serve as the role of ground truth Y in Feature CP.

De nition 3 (Surrogate feature) Consider a trained neural network $\hat{} = \hat{g} \circ \hat{f}$ where $\circ$ denotes the composition operator. For a sample (X, Y), we de ne $v = \hat{f}(X)$ to be the trained feature. Besides, we de ne the surrogate feature to be any feature v such that $\hat{g}(v) = Y$.

In contrast to commonly adopted regression or classi cation scenarios where the label is unidimensional, the dimensionality of features could be much larger. We thus de ne a corresponding non-conformity score based on the surrogate feature as follows:

$$s(X, Y, \hat{g} \circ \hat{f}) = \inf_{v \in \{v : \hat{g}(v) = Y\}} \|v - \hat{f}(X)\|. \tag{3}$$

[1] We use $_u$ to represent a Dirac Delta function (distribution) at point u.

It is usually complicated to calculate the score in Equation 3 due to the in mum operator. Therefore, we design Algorithm 2 to calculate an upper bound of the non-conformity score. Although the exact in mum is hard to achieve in practice, we can apply gradient descent starting from the trained feature $\hat{v}$ to nd a surrogate feature v around it. In order to demonstrate the reasonability of this algorithm, we analyze the non-conformity score distribution with realistic data in Appendix B.9.

4.2 Band Estimation and Band Detection

Utilizing the non-conformity score derived in Section 4.1, one could derive a con dence band in the feature space. In this section, we mainly focus on how to transfer the con dence band in feature space to the output space, i.e., calculating the set

$$\{\hat{g}(v) : \|v - \hat{v}\| \le Q_1\}, \tag{4}$$

where $\hat{v}$ is the trained feature, $\hat{g}$ is the prediction head, and Q_1 is derived based on the calibration set (even though slightly different, we refer to step 4 in Algorithm 1 for the notion of Q_1; a formal discussion of it is deferred to Algorithm 3).

Algorithm 2 Non-conformity Score

Require: Data point (X, Y), trained predictor $\hat{g} \circ \hat{f}(\cdot)$, step size α, number of steps M;

1: $u \leftarrow \hat{f}(X)$;
2: $m \leftarrow 0$;
3: **while** $m < M$ **do**
4: $\quad u \leftarrow u - \alpha \frac{\partial \|\hat{g}(u) - Y\|^2}{\partial u}$;
5: $\quad m \leftarrow m + 1$;
6: **end while**

Ensure: $s(X, Y; \hat{g} \circ \hat{f}) = \|u - \hat{f}(X)\|$.

Since the prediction head $\hat{g}$ is usually highly non-linear, the exact con dence band is hard to represent explicitly. Consequently, we provide two approaches: Band Estimation which aims at estimating the upper bound of the con dence band, and Band Detection which aims at identifying whether a response falls inside the con dence interval. We next crystallize the two methods.

Band Estimation. We model the Band Estimation problem as a perturbation analysis one, where we regard Equation (4) as a perturbation of the trained feature $\hat{v}$, and analyze the output bounds of prediction head $\hat{g}$. In this work, we apply linear relaxation based perturbation analysis (LiPRA) (Xu et al., 2020) to tackle this problem under deep neural network regimes. LiPRA transforms the certi cation problem as a linear programming problem, and solves it accordingly. The relaxation would result in a relatively looser interval than the actual band, so this method would give an upper bound estimation of the exact band length.

Band Detection. Band Estimation could potentially end up with loose inference results. Typically, we are only interested in determining whether a point Y' is in the con dence band $C(X')$ for a test sample X'. To achieve this goal, we rst apply Algorithm 2 using data point (X', Y'), which returns a non-conformity score V'. We then test whether the score V' is smaller than quantile Q_1 on the calibration set (see Equation (4)). If so, we deduce that $Y' \in C(X')$ (or vice versa if not).

4.3 Feature Conformal Prediction

Based on the above discussion, we summarize Feature CP in Algorithm 3.[2] Different from vanilla CP (see Algorithm 1), Feature CP uses a different non-conformity score based on surrogate features, and we need an additional Band Estimation or Band Detection (step 5) to transfer the band from feature space to output space.

We then discuss two intriguing strengths of Feature CP. First, the proposed technique is universal and could improve other advanced adaptive conformal inference techniques utilizing the inductive bias of learned feature space. Speci cally, we propose Feature CQR with insights from CQR (Romano et al., 2019), a prominent adaptive conformal prediction method with remarkable performance, to demonstrate the universality of our technique. We relegate related algorithmic details to Section B.4. Second, although methods such as CQR require specialized training criteria (quantile regression) for the predictive models, Feature CP could be directly applied to any given pretrained model and

[2] We here show the Feature CP algorithm with Band Estimation. We defer the practice details of using Band Detection in step 5 of Algorithm 3 to Appendix.

Algorithm 3 Feature Conformal Prediction

Require: Level α, dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i \in \mathcal{I}}$, test point X^0;

1: Randomly split the dataset $\mathcal{D}$ into a training fold $\mathcal{D}_{tr}$, $(X_i, Y_i)_{i \in \mathcal{I}_{tr}}$ together with a calibration fold $\mathcal{D}_{ca}$, $(X_i, Y_i)_{i \in \mathcal{I}_{ca}}$;
2: Train a base machine learning model $\hat{f}(\cdot)$ using $\mathcal{D}_{tr}$ to estimate the response Y;
3: For each $i \in \mathcal{I}_{ca}$, calculate the non-conformity score V_i based on $\hat{f}$ Algorithm 2;
4: Calculate the $(1-\alpha)$-th quantile $Q_{1-\alpha}$ of the distribution $\frac{1}{|\mathcal{I}_{ca}|+1} \sum_{i \in \mathcal{I}_{ca}} V_i + \infty$;
5: Apply Band Estimation on test data feature $\hat{f}(X^0)$ with perturbation $Q_{1-\alpha}$ and prediction head $\hat{g}$, which returns $C_{1-\alpha}^{fcp}(X)$;

Ensure: $C_{1-\alpha}^{fcp}(X)$.

could still give meaningful adaptive interval estimates. This trait facilitates the usage of our method with large pretrained models, which is common in modern language and vision tasks.

4.4 Theoretical Guarantee

This section presents theoretical guarantees for Feature CP regarding coverage (effectiveness) and band length (efficiency). We provide an informal statement of the theorem below and defer the complete details to Appendix A.1.

Theorem 4 (Informal Theorem on the Efficiency of Feature CP) Under mild assumptions, if the following cubic conditions hold:

1. Length Preservation. Feature CP does not cost much loss in feature space.
2. Expansion. The Band Estimation operator expands the differences between individual length and their quantiles.
3. Quantile Stability. The band length is stable in both the feature space and the output space for a given calibration set.

then Feature CP outperforms vanilla CP in terms of average band length.

The intuition of Theorem 4 is as follows: Firstly, Feature CP and Vanilla CP take quantile operations in different spaces, and the Expansion condition guarantees that the quantile step costs less in Feature CP. However, there may be an efficiency loss when transferring the band from feature space to output space. Fortunately, it is controllable under the Length Preservation condition. The final Quantile Stability condition ensures that the band is generalizable from the calibration fold to the test samples. We provide the detailed theorem in Appendix A.1 and empirically validate the cubic conditions in Appendix B.2.

5 Experiments

We conduct experiments on synthetic and real-world datasets, mainly to show that Feature CP is (a) effective, i.e., it could return valid confidence bands with empirical coverage larger than $1-\alpha$; (b) efficient, i.e., it could return shorter confidence bands than vanilla CP.

5.1 Setup

Datasets. We consider both synthetic datasets and real-world datasets, including (a) realistic uni-dimensional target datasets: five datasets from UCI machine learning repository (Asuncion, 2007): physicochemical properties of protein tertiary structure (bio), bike sharing (bike), community and crimes (community) and Facebook comment volume variants one and two (facebook 1/2), five datasets from other sources: blog feedback (blog) (Buza, 2014), Tennessee's student teacher achievement ratio (star) (Achilles et al., 2008), and medical expenditure panel survey (meps19–21) (Cohen et al., 2009); (b) synthetic multi-dimensional target dataset $Y = WX + \epsilon$, where $X \in [0,1]^{100}$; $Y \in \mathbb{R}^{10}$,

Figure 2: Performance on tasks with one-dimensional target (= 0:1). Left: Empirical coverage. Right: Con dence interval length, where smaller value is better. The proposed Feature CP and Feature CQR could consistently achieve shorter bands while maintaining a good coverage performance.

Table 1: Performance of both methods on multi-dimensional regression benchmarks (= 0:1), where "w-length" denotes the weighted con dence band length.

DATASET	SYNTHETIC		CITYSCAPES		
METHOD	COVERAGE	LENGTH	COVERAGE	LENGTH	W-LENGTH
BASELINE	89:91 1:03	0:401 0:01	91:41 0:51	40:15 0:02	40:15 0:02
FEATURE CP	90:13 0:59	0.373 0:05	90:77 0:91	1.032 0:01	0.906 0:01

follows the standard Gaussian distribution, and is a xed randomly generated matrix; and (c) real-world semantic segmentation dataset: Cityscapes (Cordts et al., 2016), where we transform the original pixel-wise classi cation problem into a high-dimensional pixel-wise regression problem. We also extend Feature CP to classi cation problems and test on the ImageNet (Deng et al., 2009) dataset. We defer more related details to Appendix B.1.

Algorithms. We compare the proposed Feature CP against the vanilla conformal baselines without further speci ed, which directly deploy conformal inference on the output space. For both methods, we use ℓ_1-type non-conformity score, namely $s(X, Y,) = \|Y - (X)\|_1$.

Evaluation. We adopt the following metrics to evaluate algorithmic empirical performance.

Empirical coverage (effectiveness) is the empirical probability that a test point falls into the predicted con dence band. A good predictive inference method should achieve empirical coverage slightly larger than $1 -$ for a given signi cance level . To calculate the coverage for Feature CP, we rst apply Band Detection on the test point (X^0, Y^0) to detect whether Y^0 is in $C_1^{fcp}(X^0)$, and then calculate its average value to obtain the empirical coverage.

Band length (ef ciency). Given the empirical coverage being larger than $1 -$, we hope the con dence band to be as short as possible. The band length should be compared under the regime of empirical coverage being larger than $1 -$, otherwise one can always set the con dence band to empty to get a zero band length. Since the explicit expression for con dence bands is intractable for the proposed algorithm, we could only derive an estimated band length via Band Estimation. Concretely, we

Figure 3: Visualization of Feature CP in image segmentation. The brightness of the pixels in the third column measures the uncertainty of Feature CP, namely the length of con dence bands. The algorithm is more uncertain in the brighter regions. For better visualization, we rescale the interval length to the range of $[0, 1]$. Feature CP is more uncertain in non-informative regions, which are object boundaries.

(a) Facebook1 (b) Synthetic (c) Cityscapes

Figure 4: Empirical coverage under different con dence levels. For a good conformal prediction method, the y-axis (i.e., empirical coverage minus $1 - \alpha$) should keep being above zero for different α. These three gures above show that Feature CP generally performs better than the baseline, in the sense that this difference is above zero most of the time.

rst use Band Estimation to estimate the con dence interval, which returns a band with explicit formulation, and then calculate the average length across each dimension.

We formulate the metrics as follows. Let $Y_{test} = (Y^{(1)}, \ldots, Y^{(d)}) \in \mathbb{R}^d$ denotes the high dimensional response and $C(X) \subseteq \mathbb{R}^d$ denotes the obtained con dence interval, with length in each dimension forming a vector $|C(X)| \in \mathbb{R}^d$. With the test set index being I_{te} and $[d] = \{1, \ldots, d\}$, we calculate the empirical coverage and band length respectively as

$$\frac{1}{|I_{te}|} \sum_{i \in I_{te}} \mathbb{I}(Y_i \in C(X_i)), \quad \frac{1}{|I_{te}|} \sum_{i \in I_{te}} \frac{1}{d} \sum_{j \in [d]} |C(X_i)|^{(j)}.$$

5.2 Results and Discussion

Effectiveness. We summarize the empirical coverage in Figure 2 (one-dimension response) and Table 1 (multi-dimension response). As Theorem 5 illustrates, the empirical coverage of Feature CP all exceeds the con dence level $1 - \alpha$, indicating that Feature CP is effective. Besides, Figure 4 demonstrates that that the effectiveness holds with different signi cance levels. For simple benchmarks such as facebook1 and synthetic data, both methods achieve similar coverage due to the simplicity; while for the harder Cityscapes segmentation task, the proposed method outperforms the baseline under many con dence levels.

Ef ciency. We summarize the con dence band in Figure 2 (one-dimension response) and Table 1 (multi-dimension response). The band length presented is an estimated version via Band Estimation.

Table 2: Feature CP outperform previous methods on large-scale classification tasks (ImageNet, $\alpha = 0.1$). We compare our results with APS (Romano et al., 2020a) and RAPS (Angelopoulos et al., 2021b). The results of baselines are taken from Angelopoulos et al. (2021b).

METHOD	ACCURACY		COVERAGE			LENGTH		
MODEL	TOP-1	TOP-5	APS	RAPS	FEATURE CP	APS	RAPS	FEATURE CP
RESNET 18	0.698	0.891	0.900	0.900	0.902 0.0023	16.2	4.43	3.82 0.18
RESNET 50	0.761	0.929	0.900	0.900	0.900 0.0047	12.3	2.57	2.14 0.03
RESNET 101	0.774	0.936	0.900	0.900	0.900 0.0025	10.7	2.25	2.17 0.01
RESNEXT 101	0.783	0.945	0.900	0.900	0.900 0.0030	19.7	2.00	1.80 0.08
SHUFFLENET	0.694	0.883	0.900	0.900	0.901 0.0019	31.9	5.05	5.06 0.04
VGG 16	0.716	0.904	0.901	0.900	0.901 0.0047	14.1	3.54	3.26 0.03

Note that Feature CP outperforms the baseline in the sense that it achieves a shorter band length and thus a more efficient algorithm.

Comparison to CQR. The techniques proposed in this paper can be generalized to other conformal prediction techniques. As an example, we propose Feature CQR which is a feature-level generalized version of CQR, whose details are deferred to Appendix B.4. We display the comparison in Figure 2, where our method consistently outperforms CQR baseline by leveraging good representation. Besides, we evaluate the group coverage performance of CQR and Feature CQR in Appendix B.5, demonstrating that Feature CQR generally outperforms CQR in terms of condition coverage. One can also generalize the other existing techniques to feature versions, e.g., Localized Conformal Prediction (Guan, 2019; Han et al., 2022).

Extension to classification tasks. The techniques in Feature CP can be generalized to classification tasks. We take the ImageNet classification dataset as an example. Table 2 shows the effectiveness of Feature CP in classification tasks. We experiment on various different architectures and use the same pretrained weights as the baselines. We provide additional details in Appendix B.7.

Ablation on the splitting point. We demonstrate that the coverage is robust to the splitting point in both small neural networks (Table 9 in Appendix) and large neural networks (Table 10 in Appendix). One can also use standard cross-validation to choose the splitting point. Since the output space is one of the special feature layers, such techniques always generalize the scope of vanilla CP.

Truthfulness. We visualize the segmentation results in Figure 3, which illustrates that Feature CP returns large bands (light region) on the non-informative regions (object boundaries) and small bands (dark region) on the informative regions. We do not show baseline visualization results since they return the same band in each dimension for each sample, and therefore does not contain much information. We also evaluate the performance with weighted band length, defined in Appendix B.1.

How does Feature CP benefit from a good deep representation? Here we provide some intuition on the success of Feature CP algorithm: we claim it is the usage of good (deep) representation that enables Feature CP to achieve better predictive inference. To validate this hypothesis, we contrast Feature CP against the baseline with an unlearned neural network (whose feature is not semantic as desired). This "random" variant of Feature CP does not outperform its vanilla counterpart with the same neural network, which confirms our hypothesis. We defer the results to Table 6 and related discussion to Appendix B.3.

More discussion. We analyze the failure (i.e., inefficient) reasons of vanilla CP in image segmentation task from the following two perspectives. Firstly, this paper aims to provide valuable coverage, namely, the confidence band should cover the ground truth for each pixel. Since vanilla CP returns the same band for different samples, the loss is pretty large such that the returned interval is large enough to cover the ground truth. Secondly, an intuitive explanation relates to our usage of ℓ_1 loss to form the non-conformity score during the training. We choose the infinity norm because reporting the total band length requires the band length in each dimension. As a result, the non-conformity score is large as long as there exists one pixel that does not fit well, contributing to an unsatisfying band for vanilla CP.

ACKNOWLEDGMENTS

This work has been partly supported by the Ministry of Science and Technology of the People's Republic of China, the 2030 Innovation Megaprojects “Program on New Generation Arti cial Intelligence” (Grant No. 2021AAA0150000). This work is also supported by a grant from the Guoqiang Institute, Tsinghua University.

REFERENCES

CM Achilles, Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word. Tennessee's student teacher achievement ratio (star) project. Harvard Dataverse, 1:2008, 2008.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quanti cation. CoRR, abs/2107.07511, 2021. URL https://arxiv.org/abs/2107.07511.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. arXiv preprint arXiv:2110.01052, 2021a.

Anastasios N. Angelopoulos, Amit P. S. Kohli, Stephen Bates, Michael I. Jordan, Jitendra Malik, Thayer Alshaabi, Srigokul Upadhyayula, and Yaniv Romano. Image-to-image regression with distribution-free uncertainty quanti cation and applications in imaging. CoRR, abs/2202.05265, 2022. URL https://arxiv.org/abs/2202.05265.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. Uncertainty sets for image classi ers using conformal prediction. 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021b. URL https://openreview.net/forum?id=eNdiU_DbM9.

Arthur U. Asuncion. Uci machine learning repository, university of california, irvine, school of information and computer sciences. 2007.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. The limits of distribution-free conditional predictive inference. Information and Inference: A Journal of the IMA, 2020.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. arXiv preprint arXiv:2202.13415, 2022.

Stephen Bates, Anastasios Nikolas Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution free, risk controlling prediction sets, 2021. URL https://arxiv.org/abs/2101.02703.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis R. Bach and David M. Blei (eds.), Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, volume 37 of JMLR Workshop and Conference Proceedings, pp. 1613–1622. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/blundell15.html.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Krisztian Buza. Feedback prediction for blogs. In *Data analysis, machine learning and knowledge discovery*, pp. 145–152. Springer, 2014.

Emmanuel J Candès, Lihua Lei, and Zhimei Ren. Conformalized survival analysis. *arXiv preprint arXiv:2103.09763*, 2021.

Yanzhi Chen, Dinghuai Zhang, Michael U. Gutmann, Aaron Courville, and Zhanxing Zhu. Neural approximate suf cient statistics for implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=SRDuJssQud .

Joel W. Cohen, Steven B. Cohen, and Jessica S. Banthin. The medical expenditure panel survey: A national information resource to support healthcare cost research and inform policy and practice. *Medical Care*, 47:S44–S50, 2009.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://doi.org/10.18653/v1/n19-1423 .

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *CoRR*, abs/2107.03342, 2021. URL https://arxiv.org/abs/2107.03342 .

Asaf Gendler, Tsui-Wei Weng, Luca Daniel, and Yaniv Romano. Adversarially robust conformal prediction. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9L1BsI4wP1H .

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. In *NeurIPS*, 2021.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017a. URL http://proceedings.mlr.press/v70/guo17a.html .

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *ArXiv*, abs/1706.04599, 2017b.

Xing Han, Ziyang Tang, Joydeep Ghosh, and Qiang Liu. Split localized conformal prediction. *CoRR*, abs/2206.13092, 2022. doi: 10.48550/arXiv.2206.13092. URL https://doi.org/10.48550/arXiv.2206.13092 .

Robert M. Haralick and Linda G. Shapiro. Image segmentation techniques. *Comput. Vis. Graph. Image Process.*, 29(1):100–132, 1985. doi: 10.1016/S0734-189X(85)90153-7. URL https://doi.org/10.1016/S0734-189X(85)90153-7 .

José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 1861–1869. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/hernandez-lobatoc15.html.

Xiaoyu Hu and Jing Lei. A distribution-free test of covariate shift using conformal prediction. *arXiv preprint arXiv: Methodology*, 2020.

Rafael Izbicki, Gilson Shimizu, and Rafael B Stern. Cd-split and hpd-split: efficient conformal regions in high dimensions. *arXiv preprint arXiv:2007.12778*, 2020a.

Rafael Izbicki, Gilson T. Shimizu, and Rafael Bassi Stern. Distribution-free conditional predictive bands using density estimators. *ArXiv*, abs/1910.05575, 2020b.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2801–2809. PMLR, 2018. URL http://proceedings.mlr.press/v80/kuleshov18a.html.

James Law. Review of "algorithmic learning in a random world by vovk, gammerman and shafer", springer, 2005, ISBN: 0-387-00152-2. *SIGACT News*, 37(4):38–40, 2006. doi: 10.1145/1189056.1189065. URL https://doi.org/10.1145/1189056.1189065.

Jing Lei, Alessandro Rinaldo, and Larry A. Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74:29–43, 2013.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018.

Lihua Lei and Emmanuel J. Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021a.

Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021b.

Yingzhen Li and Yarin Gal. Dropout inference in bayesian neural networks with alpha-divergences. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2052–2061. PMLR, 2017. URL http://proceedings.mlr.press/v70/li17a.html.

Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *CoRR*, abs/2001.05566, 2020. URL https://arxiv.org/abs/2001.05566.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Ann Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *NeurIPS*, 2021.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 38–41. Computer Vision Foundation / IEEE, 2019. URL http://openaccess.thecvf.com/content_CVPRW_2019/html/Uncertainty_and_Robustness_in_Deep_Visual_Learning/Nixon_Measuring_Calibration_in_Deep_Learning_CVPRW_2019_paper.html.

Ilia Nouretdinov, Sergi G. Costafreda, Alexander Gammerman, Alexey Ya. Chervonenkis, Vladimir Vovk, Vladimir Vapnik, and Cynthia H. Y. Fu. Machine learning classification with confidence: Application of transductive conformal predictors to mri-based diagnostic and prognostic markers in depression. *NeuroImage*, 56(2):809–813, 2011. doi: 10.1016/j.neuroimage.2010.05.023. URL https://doi.org/10.1016/j.neuroimage.2010.05.023.

Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Regression conformal prediction with nearest neighbours. J. Artif. Intell. Res., 40:815–840, 2011. URL http://jair.org/papers/paper3198.html.

Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147, 2016.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. In UAI, 2021.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp. 3538–3548, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. arXiv: Methodology, 2020b.

N. Senthilkumaran and Reghunadhan Rajesh. Image segmentation - A survey of soft computing approaches. In ARTCom 2009, International Conference on Advances in Recent Technologies in Communication and Computing, Kottayam, Kerala, India, 27-28 October, 2009, pp. 844–846. IEEE Computer Society, 2009. doi: 10.1109/ARTCom.2009.219. URL https://doi.org/10.1109/ARTCom.2009.219.

Matteo Sesia and Emmanuel J Candès. A comparison of some conformal quantile regression methods. Stat, 9(1):e261, 2020.

Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, pp. 6304–6315, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/31b3b31a1c2f8a370206f111127c0dbd-Abstract.html.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. J. Mach. Learn. Res., 9: 371–421, 2008. URL https://dl.acm.org/citation.cfm?id=1390693.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. Nat., 550(7676):354–359, 2017. doi: 10.1038/nature24270. URL https://doi.org/10.1038/nature24270.

Ralph C. Smith. Uncertainty Quantification - Theory, Implementation, and Applications, volume 12 of Computational science and engineering. SIAM, 2014. ISBN 978-1-611973-21-1. URL https://bookstore.siam.org/cs12/.

David Stutz, Krishnamurthy Dvijotham, Ali Taylan Cemgil, and A. Doucet. Learning optimal conformal classifiers. ArXiv, abs/2110.09192, 2021.

Timothy John Sullivan. Introduction to uncertainty quantification, volume 63. Springer, 2015.

Jiaye Teng, Zeren Tan, and Yang Yuan. T-SCI: A two-stage conformal inference algorithm with guaranteed coverage for cox-mlp. In Marina Meila and Tong Zhang (eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event volume 139 of Proceedings of Machine Learning Research, pp. 10203–10213. PMLR, 2021. URL http://proceedings.mlr.press/v139/teng21a.html .

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pp. 2526–2536, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html .

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. Algorithmic learning in a random world. Springer Science & Business Media, 2005.

Chen Xu and Yao Xie. Conformal prediction interval for dynamic time-series. ICML, 2021.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0cbc5671ae26f67871cb914d81ef8fc1-Abstract.html .

Yachong Yang and Arun Kumar Kuchibhotla. Finite-sample efficient conformal prediction. arXiv preprint arXiv:2104.13871, 2021.

Appendix

Section A provides the complete proofs, and Section B.1 provides experiment details.

A THEORETICAL PROOFS

We first show the formal version of Theorem 4 in Section A.1, and show its proof in Section A.2. Theorem 5 and Theorem 6 shows the effectiveness (empirical coverage) and the efficiency (band length) in Feature CP.

We additionally provide Theorem 9 (see Section A.3) and Theorem 15 (see Section A.4) to better validate our theorem, in terms of the length variance and convergence rate.

A.1 THEORETICAL GUARANTEE

This section provides theoretical guarantees for Feature CP regarding coverage (effectiveness) and band length (efficiency), starting from additional notations.

Notations. Let P denote the population distribution. Let $\mathcal{D}_{ca} \sim P^n$ denote the calibration set with sample size n and sample index I_{ca}, where we overload the notation P^n to denote the distribution of a set with samples drawn from distribution P. Given the model $\hat{f}$ with feature extractor $\hat{f}^1$ and prediction head $\hat{f}^2$, we assume $\hat{f}^2$ is continuous. We also overload the notation $Q_{1-\alpha}(V)$ to denote the $(1-\alpha)$-quantile of the set $V \cup \{+\infty\}$. Besides, let $M[\cdot]$ denote the mean of a set, and a set minus a real number denote the broadcast operation.

Vanilla CP. Let $V^o_{\mathcal{D}_{ca}} = \{v^o_i\}_{i \in I_{ca}}$ denote the individual length in the output space for vanilla CP, given the calibration set $\mathcal{D}_{ca}$. Concretely, $v^o_i = 2|y_i - \hat{y}_i|$ where y_i denotes the true response of sample i and $\hat{y}_i$ denotes the corresponding prediction. Since vanilla CP returns band length with $1-\alpha$ quantile of non-conformity score, the resulting average band length is derived by $Q_{1-\alpha}(V^o_{\mathcal{D}_{ca}})$.

Feature CP. Let $V^f_{\mathcal{D}_{ca}} = \{v^f_i\}_{i \in I_{ca}}$ be the individual length (or diameter in high dimensional cases) in the feature space for Feature CP, given the calibration set $\mathcal{D}_{ca}$. To characterize the band length in the output space, we define $H(v; X)$ as the individual length on sample X in the output space, given the length v in the feature space, i.e., $H(v; X)$ represents the length of the set $\{g(u) \in \mathbb{R} : \|u - \hat{f}^1(X)\| \le v/2\}$. Due to the continuity assumption on function g, the above set is always simply-connected. We here omit the dependency of prediction head $\hat{f}^2$ and H for simplicity. The resulting band length in Feature CP is denoted by $\mathbb{E}_{(X^0,Y^0)\sim P} H(Q_{1-\alpha}(V^f_{\mathcal{D}_{ca}}); X^0)$. Without abuse of notations, operating H on a dataset $(g, H(V^f_{\mathcal{D}_{ca}}; \mathcal{D}_{ca}))$ means operating H on each data point $(v^f_i; X_i)$ in the set.

Coverage guarantee. We next provide theoretical guarantees for Feature CP in Theorem 5, which informally shows that under Assumption 1, the confidence band returned by Algorithm 3 is valid, meaning that the coverage is provably larger than $1-\alpha$. We defer the whole proof to Appendix A.2.

Theorem 5 (theoretical guarantee for Feature CP). Under Assumption 1, for any $\alpha > 0$, the confidence band returned by Algorithm 3 satisfies:

$$P(Y^0 \in C^{fcp}_{1-\alpha}(X^0)) \ge 1 - \alpha,$$

where the probability is taken over the calibration fold and the testing point (X^0, Y^0).

Length (efficiency) guarantee. We next show in Theorem 6 that Feature CP is provably more efficient than the vanilla CP.

Theorem 6 (Feature CP is provably more efficient). Assume that the non-conformity score is in norm-type. For the operator H, we assume a Holder assumption that there exist $\delta > 0, L > 0$ such that $|H(v; X) - H(u; X)| \le L|v - u|$ for all X. Besides, we assume that there exists $\delta_0, c > 0$, such that the feature space satisfies the following cubic conditions:

1. Length Preservation. Feature CP does not cost much loss in feature space in a quantile manner, namely $\mathbb{E}_{\mathcal{D}\sim P^n} Q_{1-\alpha}(H(V^f_{\mathcal{D}}; \mathcal{D})) < \mathbb{E}_{\mathcal{D}\sim P^n} Q_{1-\alpha}(V^o_{\mathcal{D}}) + \epsilon$.

2. Expansion. The operator $H(v; X)$ expands the differences between individual length and their quantiles, namely, $L E_{D \sim P^n} M|Q_1(V_D^f) - V_D^f| < E_{D \sim P^n} M[Q_1(H(V_D^f; D)) - H(V_D^f; D)] \leq 2\max\{L, 1\}(c/\sqrt{n})^{\min\{\alpha, 1\}}$.
3. Quantile Stability. Given a calibration set D_{ca}, the quantile of the band length is stable in both feature space and output space, namely $E_{D \sim P^n}|Q_1(V_D^f) - Q_1(V_{D_{ca}}^f)| \leq \frac{c}{\sqrt{n}}$ and $E_{D \sim P^n}|Q_1(V_D^o) - Q_1(V_{D_{ca}}^o)| \leq \frac{c}{\sqrt{n}}$.

Then Feature CP provably outperforms vanilla CP in terms of average band length, namely,

$$E H(Q_1(V_{D_{ca}}^f); X^0) < Q_1(V_{D_{ca}}^o);$$

where the expectation is taken over the calibration fold and the testing point (X^0, Y^0).

The cubic conditions used in Theorem 6 sketch the properties of feature space from different aspects. The first condition implies that the feature space is efficient for each individual, which holds when the band is generally not too large. The second condition is the core of the proof, which informally assumes that the difference between quantile and each individual is smaller in feature space. Therefore, conducting quantile operation would not harm the effectiveness (namely, step 4 in Algorithm 1 and step 4 in Algorithm 3), leading to the efficiency of Feature CP. The last condition helps generalize the results from the calibration set to the test set.

Proof of Theorem 6. We start the proof with Assumption 2, which claims that

$$L E_D M|Q_1(V_D^f) - V_D^f| < E_D M\left[Q_1(H(V_D^f; D)) - H(V_D^f; D)\right] \leq 2\max\{L, 1\}(c/\sqrt{n})^{\min\{\alpha, 1\}}.$$

We rewrite it as

$$E_D M H(V_D^f; D) < E_D Q_1(H(V_D^f; D)) + 2\max\{L, 1\}(c/\sqrt{n})^{\min\{\alpha, 1\}} - L E_D M|Q_1(V_D^f) - V_D^f|.$$

Due to Holder condition, we have that $H(Q_1(V_D^f); D) < M(H(V_D^f; D)) + L M|Q_1(V_D^f) - V_D^f|$, therefore

$$E_D M\left[H(Q_1(V_D^f); D)\right] < E_D Q_1(H(V_D^f; D)) + 2\max\{1, L\}[c/\sqrt{n}]^{\min\{1, \alpha\}}.$$

Therefore, due to assumption 1, we have that

$$E_D M H(Q_1(V_D^f); D) < E_D Q_1(V_D^o) + 2\max\{1, L\}[c/\sqrt{n}]^{\min\{1, \alpha\}}.$$

Besides, according to the quantile stability assumption, we have that $E|M H(Q_1(V_D^f); D) - M H(Q_1(V_D^f); D)| \leq L[c/\sqrt{n}]$, and $E_D|Q_1(V_D^o) - Q_1(V_D^o)| \leq c/\sqrt{n}$. Therefore,

$$\begin{aligned} & E H(Q_1(V_{D_{ca}}^f); X^0) \\ = & E_D M H(Q_1(V_{D_{ca}}^f); D) \\ < & Q_1(V_{D_{ca}}^o) + 2\max\{1, L\}[c/\sqrt{n}]^{\min\{1, \alpha\}} + L[c/\sqrt{n}] + c/\sqrt{n} \\ < & Q_1(V_{D_{ca}}^o). \end{aligned}$$

□

A.1.1 Example for Theorem 6

This section provides an example for Theorem 6. The key information is that Feature CP loses less efficiency when conducting the quantile step.

Table 3: A concrete example for the comparison between Feature CP and CP. Let IL_f denote the individual length in the feature and output space. $Q(\cdot)$ denote the quantile operator, and $H(\cdot)$ denote the operator that calculates output space length given the feature space length. We remark that the average band length returned by Feature CP (3.1) outperforms that of vanilla CP (4.0).

METHOD	VANILLA CP		FEATURE CP			
SAMPLE	IL_o	$Q(IL_o)$	IL_f	$H(IL_f)$	$Q(IL_f)$	$H(Q(IL_f))$
A	1.0	4.0	1.1	1.2	1.3	1.4
B	2.0	4.0	1.2	2.1	1.3	2.3
C	3.0	4.0	1.1	2.8	1.3	3.1
D	4.0	4.0	1.3	3.8	1.3	3.8
E	5.0	4.0	1.6	5.2	1.3	4.9
QUANTILE	4.0	/	1.3	/	/	/
AVERAGE	/	4.0	/	/	/	3.1

Assume the dataset has five samples labeled A, B, C, D, and E. When directly applying vanilla CP leads to individual length in the output space IL_o as $1; 2; 3; 4; 5$, respectively. By taking 80% quantile (namely, $= 0:2$), the final confidence band returned by vanilla CP, $Q(IL_o)$ would be $Q_{0:8}(f 1; 2; 3; 4; 5g) = 4$. Note that for any sample, the returned band length would be 4, and the final average band length is 4.

We next consider Feature CP. We assume that the individual length in the feature space, IL_f is $1:1; 1:2; 1:1; 1:3; 1:6$, respectively. Due to the expansion condition (cubic condition #2), the difference between IL_f and $Q(IL_f)$ is smaller than that between IL_o and $Q(IL_o)$. Therefore, the quantile step costs less in Feature CP. Since IL_f is close to $Q(IL_f)$, their corresponding output length $H(IL_f)$, $H(Q(IL_f))$ are also close. Besides, to link conformal prediction and vanilla CP, the length preservation condition (cubic condition #1) ensures that IL_o is close to $H(IL_f)$. Therefore, the final average length $H(Q(L_f))$ is close to the average length IL_o, which is better than $Q(IL_o)$ Finally, the quantile stability condition (cubic condition #3) generalizes the results from the calibration set to the test set.

A.2 PROOF OF THEOREM 5

Theorem 5 (theoretical guarantee for Feature CP) Under Assumption 1, for any > 0, the confidence band returned by Algorithm 3 satisfies:

$$P(Y^0 2 C_1^{fcp}(X^0)) \quad 1 \quad ;$$

where the probability is taken over the calibration fold and the testing point $(X^0; Y^0)$.

Proof of Theorem 5. The key to the proof is to derive the exchangeability of the non-conformity score, given that the data in the calibration fold and test fold are exchangeable (see Assumption 1).

For ease of notations, we denote the data points in the calibration fold and the test fold as $D^0 = f(X_i; Y_i)g_{i2[m]}$, where m denotes the number of data points in both calibration fold and test fold. By Assumption 1, the data points in D^0 are exchangeable.

The proof can be split into three parts. The first step is to show that for any function independent of D^0, $h(X_i; Y_i)$ are exchangeable. The second step is to show that the proposed score function satisfies the above requirements. And the third step is to show the theoretical guarantee based on the exchangeability of the non-conformity score.

We next prove the first step: for any given function $h : X \quad Y ! \quad R$ that is independent of data points in D^0, we have that $h(X_i; Y_i)$ are exchangeable. Specifically, its CDF F_v and its perturbation CDF

F_v is the same, given the training fold $\mathcal{D}_{tr}$.

$$\begin{aligned}
&F_v(u_1, \dots, u_n \mid \mathcal{D}_{tr}) \\
&= P(h(X_1, Y_1) \le u_1, \dots, h(X_n, Y_n) \le u_n \mid \mathcal{D}_{tr}) \\
&= P((X_1, Y_1) \in C_{h,1}(u_1), \dots, (X_n, Y_n) \in C_{h,1}(u_n) \mid \mathcal{D}_{tr}) \\
&= P((X_{(1)}, Y_{(1)}) \in C_{h,1}(u_1), \dots, (X_{(n)}, Y_{(n)}) \in C_{h,1}(u_n) \mid \mathcal{D}_{tr}) \\
&= P(h(X_{(1)}, Y_{(1)}) \le u_1, \dots, h(X_{(n)}, Y_{(n)}) \le u_n \mid \mathcal{D}_{tr}) \\
&= F_v(u_1, \dots, u_n \mid \mathcal{D}_{tr}),
\end{aligned}$$

where $\cdot$ denotes a random perturbation, and $C_{h,1}(u) = \{(X, Y) : h(X, Y) \le u\}$.

The second step is to show that the proposed non-conformity score function (See Equation (3) and Algorithm 3) is independent of the dataset $\mathcal{D}^0$. To show that, we note that the proposed score function s in Equation (3) (we rewrite it in Equation (5)) is totally independent of dataset $\mathcal{D}^0$, in that we only use the information of $\hat{f}$ and $\hat{g}$ which is dependent on the training fold $\mathcal{D}_{tr}$ instead of $\mathcal{D}^0$.

$$s(X, Y; \hat{g}, \hat{f}) = \inf_{v \in \{v : \hat{g}(v) = Y\}} \|v - \hat{f}(X)\|. \tag{5}$$

Besides, note that when calculating the non-conformity score in Algorithm 3 for each testing data/calibration data, we do not access any information on the calibration folds for any other points. Therefore, the score function does not depend on the calibration fold or test fold. We finally remark that here we always state that the score function s does not depend on the calibration fold or test fold, but its realizations $s(X, Y; \hat{g}, \hat{f})$ can depend on the two folds, if $(X, Y) \in \mathcal{D}^0$. This does not contrast with the requirement in the first step.

Therefore, combining the two steps leads to a conclusion that the non-conformity scores are on $\mathcal{D}^0$ are exchangeable. Finally, following Lemma 1 in Tibshirani et al. (2019), the theoretical guarantee holds under the exchangeability of non-conformity scores. □

A.3 Length Variance Guarantee

The next Theorem 9 demonstrates that the length returned by Feature CP would be individually different. Specifically, the variance for the length is lower bounded by a constant. The essential intuition is that, for a non-linear function g, the feature bands with the same length return different bands in output space. Before expressing the theorem, we first introduce a formal notation of length and other necessary assumptions. For ease of discussion, we define in Definition 7 a type of band length slightly different from the previous analysis. We assume $Y \in \mathbb{R}$ below, albeit our analysis can be directly extended to high-dimensional cases.

Definition 7 (band length) For a given feature v and any perturbation $\tilde{v} \in C_f(v) = \{\tilde{v} : \|\tilde{v} - v\| \le Q\}$ in the feature band, we define the band length in the output space $L_o(v)$ as the maximum distance between predictor $g(v)$ and $g(\tilde{v})$, namely

$$L_o(v) \triangleq \max_{\tilde{v} \in C_f(v)} |g(\tilde{v}) - g(v)|.$$

Besides, we require Assumption 8, which is about the smoothness of the prediction head g.

Assumption 8. Assume that the prediction head g is second order derivative and M-smooth, namely, $\|\nabla^2 g(u)\| \le M$ for all feasible u.

The following Theorem 9 indicates that the variance of the band length is lower bounded, meaning that the bands given by Feature CP are individually different.

Theorem 9. Under Assumption 8, if the band on the feature space is with radius Q, then the variance of band length on the output space satisfies:

$$\mathbb{E}[L_o - \mathbb{E}L_o]^2 = Q^2 \cdot \mathbb{E}[\|\nabla g(v)\| - \mathbb{E}\|\nabla g(v)\|]^2 \le MQ\,\mathbb{E}\|\nabla g(v)\|.$$

From Theorem 9, the variance of the band length has a non-vacuous lower bound if

$$\mathbb{E}[\|\nabla g(v)\| - \mathbb{E}\|\nabla g(v)\|]^2 > MQ \cdot \mathbb{E}\|\nabla g(v)\|. \tag{6}$$

We next discuss the condition for Equation (6). For a linear function g, note that $\mathbb{E}[\|r\, g(v)\| - \mathbb{E}\|r\, g(v)\|] = 0$ and $M = 0$, thus does not meet Equation (6). But for any other non-linear function g, we at least have $\mathbb{E}[\|r\, g(v)\| - \mathbb{E}\|r\, g(v)\|]^2 > 0$ and $M > 0$, and therefore there exists a term Q such that Equation (6) holds. Hence, the band length in feature space must be individually different for a non-linear function g and a small band length Q.

Proof of Theorem 9. We revisit the notation in the main text, where $v = f(X)$ denotes the feature, and $C_f(v) = \{v' : \|v' - v\| \le Q\}$ denotes the conformal band returned in feature space. By Taylor Expansion, for any given $v' \in C_f(v)$, there exists a v^0 such that

$$g(v') - g(v) = r\, g(v)(v' - v) + 1/2 (v' - v)^\top r^2 g(v^0)(v' - v).$$

Due to Assumption 8, $\|r^2 g(v^0)\| \le M$. Therefore, for any $v' \in C_f(v)$

$$|1/2 (v' - v)^\top r^2 g(v^0)(v' - v)| \le \frac{1}{2} M Q^2.$$

On the one hand, by Cauchy Schwarz inequality, we have

$$L_o = \max_{v'} |g(v') - g(v)| \le \|r\, g(v)\| Q + \frac{1}{2} M Q^2.$$

On the other hand, by setting $v' = v + Q\, r\, g(v)/\|r\, g(v)\|$, we have that

$$L_o = \max_{v'} |g(v') - g(v)| \ge |g(v + Q\, r\, g(v)/\|r\, g(v)\|) - g(v)| = Q\|r\, g(v)\| - 1/2 M Q^2.$$

Therefore, we have that

$$|L_o - Q\|r\, g(v)\|| \le 1/2 M Q^2.$$

We finally show the variance of the length, where the randomness is taken over the data,

$$\begin{aligned}
&\mathbb{E}\,[L_o - \mathbb{E}L_o]^2 \\
&= \mathbb{E}[Q\|r\, g(v)\| - \mathbb{E}Q\|r\, g(v)\|] + [L_o - Q\|r\, g(v)\|] - \mathbb{E}[L_o - Q\|r\, g(v)\|]]^2 \\
&= \mathbb{E}[Q\|r\, g(v)\| - \mathbb{E}Q\|r\, g(v)\|]^2 + \mathbb{E}[[L_o - Q\|r\, g(v)\|] - \mathbb{E}[L_o - Q\|r\, g(v)\|]]^2 \\
&\quad + 2\,\mathbb{E}[Q\|r\, g(v)\| - \mathbb{E}Q\|r\, g(v)\|][(L_o - Q\|r\, g(v)\|) - \mathbb{E}(L_o - Q\|r\, g(v)\|)] \\
&\le Q^2 \mathbb{E}[\|r\, g(v)\| - \mathbb{E}\|r\, g(v)\|]^2 \\
&\quad + 2Q\mathbb{E}|[\|r\, g(v)\| - \mathbb{E}\|r\, g(v)\|]|\,|[(L_o - Q\|r\, g(v)\|) - \mathbb{E}(L_o - Q\|r\, g(v)\|)]| \\
&\le Q^2 \mathbb{E}[\|r\, g(v)\| - \mathbb{E}\|r\, g(v)\|]^2 + M Q^3 \mathbb{E}|[\|r\, g(v)\| - \mathbb{E}\|r\, g(v)\|]|.
\end{aligned}$$

Besides, note that $\mathbb{E}|[\|r\, g(v)\| - \mathbb{E}\|r\, g(v)\|]| \le \mathbb{E}\|r\, g(v)\|$. Therefore, we have that

$$\mathbb{E}\,[L_o - \mathbb{E}L_o]^2 / Q^2 \le \mathbb{E}[\|r\, g(v)\| - \mathbb{E}\|r\, g(v)\|]^2 + M Q\, \mathbb{E}\|r\, g(v)\|.$$

□

A.4 Theoretical Convergence Rate

In this section, we prove the theoretical convergence rate for the width. Specifically, we derive that when the number of samples in the calibration fold goes to infinity, the width for the testing point converges to a fixed value. Before we introduce the main theorem, we introduce some necessary definitions. Without further clarification, we follow the notations in the main text.

Definition 10 (Precise Band). We define the precise band as

$$C_1^{pre} = \{g(v) : \|v' - v\| \le Q_1 \quad \forall \|v' - v\| \le Q\}. \tag{7}$$

Definition 11 (Precise Exact Band). We define the exact precise band as

$$C_1^{pre} = \{g(v) : \|v' - v\| \le Q_1 \quad \forall \|v' - v\| \le Q\}, \tag{8}$$

where Q_1 denotes the exact value such that

$$P(\exists v' : \|v' - v\| \le Q_1 \;; g(v') = y) = 1 - \alpha. \tag{9}$$

Our goal is to prove that the band length (volume) of C_1^{pre} (denoted by $V(C_1^{pre})$) converges to $V(C_1^{pre})$. We assume that the prediction head and the quantile function are both Lipschitz in Assumption 12 and Assumption 13.

Assumption 12 (Lipschitz for Prediction Head) Assume that for any v, v', we have

$$\|g(v) - g(v')\| \le L_1 \|v - v'\|.$$

Assumption 13 (Lipschitz for Inverse Quantile Function) Denote the quantile function as

$$\text{Quantile}(Q_u) = P(\exists v : \|v - \hat{v}\| \le Q_u, g(v) = y) = u.$$

We assume that its inverse function L_2-Lipschitz, that is to say,

$$\|\text{Quantile}^{-1}(u) - \text{Quantile}^{-1}(u')\| \le L_2 |u - u'|.$$

Besides, we assume that the region of C_1^{pre} has benign blow-up.

Assumption 14 (Benign Blow-up) Assume that C_1^{pre} has benign blow-up, that is to say, for the blow-up set $C_1^{pre}(\epsilon) = \{v : \exists u \in C_1^{pre}, \|u - v\| \le \epsilon\}$, we have

$$\|V(C_1^{pre}(\epsilon)) - V(C_1^{pre})\| \le c\epsilon,$$

where c denotes a constant independent of ϵ.

In the one-dimensional case $v \in \mathbb{R}$, Assumption 14 easily holds. For the high-dimensional cases, such a bound usually requires that c depends on the dimension d.

Theorem 15 (Convergence Rate) Assume that the non-conformity scores in the calibration fold have no ties. Under Assumption 12, Assumption 13 and Assumption 14, we have that

$$\|V(C_1^{pre}) - V(C_1^{pre})\| \le cL_1L_2\frac{1}{n}.$$

Proof. Firstly, as derived in Romano et al. (2019), when the non-conformity score in the calibration fold has no ties (the probability is zero), we have

$$P(\exists v : \|v - \hat{v}\| \le Q_1, g(v) = y) \in (1 - \alpha, 1 - \alpha + 1/n), \tag{10}$$

where $v, \hat{v}, Q_1$ denotes the surrogate feature, the trained feature, and the quantile value in Algorithm 3, respectively.

By Assumption 13 that the inverse quantile function is L_2-Lipschitz around $1-\alpha$, we have

$$\|Q_1 - Q_{1-\alpha}\| \le L_2\frac{1}{n}.$$

Therefore, for any $u \in C_1^{pre}$, there exists $u' \in C_1^{pre}$ such that

$$\|u - u'\| \le L_2\frac{1}{n}, \quad \|g(v) - g(v')\| \le L_1\|v - v'\| \le L_1L_2\frac{1}{n}. \tag{11}$$

We note that bounding $\|v - v'\|$ requires that the region of v, v' are both balls, and therefore one can select v' as the point with the smallest distance to v. Since the region of C_1^{pre} has benign blow-up, we have that

$$V(C_1^{pre}) \subseteq V(C_1^{pre}) + cL_1L_2\frac{1}{n}.$$

Besides, the following equation naturally holds due to Equation (10).

$$V(C_1^{pre}) \supseteq V(C_1^{pre}).$$

Therefore, we conclude with the following inequality,

$$\|V(C_1^{pre}) - V(C_1^{pre})\| \le cL_1L_2\frac{1}{n}.$$

Therefore, as the sample size in the calibration fold goes to infinity, the length of the trained band converges to $V(C_1^{pre})$.

□

Figure 5: The model architecture of the uni-dimensional and synthetic multi-dimensional target regression experiments. The dropout layers are omitted.

Figure 6: The model architecture of the semantic segmentation experiment.

B EXPERIMENTAL DETAILS

Section B.1 introduces the omitted experimental details. Section B.2 provide experimental evidence to validate cubic conditions. Section B.3 shows that Feature CP performs similarly to vanilla CP for untrained neural networks, validating that Feature CP works due to semantic information trained in feature space. Section B.4 introduces Feature CQR which applies feature-level techniques on CQR and Section B.5 reports the corresponding group coverage. Finally, Section B.9 provides other additional experiments omitted in the main text.

B.1 EXPERIMENTAL DETAILS

Model Architecture. The model architecture of the uni-dimensional and synthetic multi-dimensional target regression task is shown in Figure 5. The feature function f and prediction head g includes two linear layers, respectively. Moreover, the model architecture of the FCN used in the semantic segmentation experiment is shown in Figure 6, which follows the of cial implementation of PyTorch. The batch normalization and dropout layers are omitted in the gure. We use the ResNet50 backbone as f and take two convolution layers as g. We select the Layer4 output of ResNet50 as our surrogate feature v.

Training protocols. In the unidimensional and synthetic dimensional target regression experiments, we randomly divide the dataset into training, calibration, and test sets with the proportion $2:2:1$. As for the semantic segmentation experiment, because the labels of the pre-divided test set are not accessible, we re-split the training, calibration, and test sets randomly on the original training set of Cityscapes. We remove the class 0 (unlabeled) from the labels during calibration and testing, and use the weighted mean square error as the training objective where the class weights are adopted from Paszke et al. (2016).

Band Estimation. In estimating the band length, we deploy Band Estimation based on the score calculated by Algorithm 2. In our experiment, we choose the number of steps M in Algorithm 2 via cross-validation.

Table 4: Comparison between CP and Feature CP.

METHOD	CP		FEATURE CP	
DATASET	COVERAGE	LENGTH	COVERAGE	LENGTH
COMMUNITY	89.82 0.95	1.99 0.09	89.62 0.83	1.99 0.19
FACEBOOK1	90.11 0.33	3.17 0.59	90.07 0.32	2.08 0.17
FACEBOOK2	89.99 0.18	2.72 0.37	89.98 0.17	1.97 0.17
MEPS19	90.51 0.21	4.25 0.21	90.55 0.17	3.58 0.35
MEPS20	89.80 0.61	4.17 0.36	89.76 0.61	3.37 0.38
MEPS21	89.92 0.54	4.67 0.30	89.94 0.54	3.93 0.54
STAR	90.30 1.17	0.41 0.02	90.35 0.99	0.41 0.03
BIO	90.27 0.26	1.97 0.01	90.20 0.29	1.87 0.04
BLOG	90.08 0.27	3.32 0.38	90.06 0.33	2.81 0.40
BIKE	89.65 0.87	1.91 0.04	89.61 0.69	1.79 0.08

Randomness. We train each model five times with different random seeds and report the mean and standard deviation value across all the runs as the experimental results (as shown in Figure 2 and Table 1).

Details of transforming segmentation classification problem into a regression task. The original semantic segmentation problem is to fit the one-hot label whose size is (C, W, H) via logistic regression, where C is the number of the classes, W, and H are the width and height of the image. We use Gaussian Blur to smooth the values in each channel. At this time, the smoothed label y ranges from 0 to 1. Then, we use the double log trick to convert the label space $[0, 1]$ to $(-\infty, +\infty)$, i.e., $y = \log(-\log(y))$. Finally, we use mean square error loss to fit y.

Definition of weighted length. We formulate the weighted length as

$$\text{weighted length} = \frac{1}{|I_{\text{test}}|} \sum_{i \in I_{\text{test}}} \sum_{j \in [d]} w_i^{(j)} |C(X_i)^{(j)}|,$$

where $w_i^{(j)}$ is the corresponding weight in each dimension. We remark that although the formulation of $w_i^{(j)}$ is usually sample-dependent, we omit the dependency of the sample and denote it by $w^{(j)}$ when the context is clear. We next show how to define $w^{(j)}$ in practice.

Generally speaking, we hope that $w^{(j)}$ is large when being informative (i.e., in non-boundary regions). Therefore, for the j-th pixel after Gaussian Blur whose value is $Y^{(j)} \in [0, 1]$, its corresponding weight is defined as

$$w^{(j)} = \frac{|2Y^{(j)} - 1|}{W} \in [0, 1],$$

where $W = \sum_j |2Y^{(j)} - 1|$ is a scaling factor.

At a colloquial level, $w^{(j)}$ is close to 1 if $Y^{(j)}$ is close to 0 or 1. In this case, $Y^{(j)}$ being close to 0 or 1 means that the pixel is far from the boundary region. Therefore, the weight indicates the degree to which a pixel is being informative (not in object boundary regions).

Calibration details. During calibration, to get the best value for the number of steps, we take a subset (one-fifth) of the calibration set as the additional validation set. We calculate the non-conformity score on the rest of the calibration set with various values of Step and then evaluate on the validation set to get the best Step whose coverage is just over $1 - \alpha$. The final trained surrogate feature v is close to the true feature because $Q(v)$ is sufficiently close to the ground truth Y. In practice, the surrogate feature after optimization satisfies $\frac{\|Q(v) - Y\|_2}{\|Y\|_2} < 1\%$.

Comparison between Feature CP and Vanilla CP. We next present the specific statistics of Figure 2 in Table 4 and Table 7.

Table 5: Validate cubic conditions.

SPACE	FEATURE SPACE		OUTPUT SPACE	
METRIC	$M\|Q_1 - V^f_{D_{CA}}\| - \|V^f_{D_{CA}}\|$		$M[Q_1 - H(V^f_{D_{CA}}) - H(V^f_{D_{CA}})]$	
COMMUNITY	0.1150	0.0290	0.8073	0.1450
FACEBOOK1	0.2491	0.0391	1.8950	0.2058
FACEBOOK2	0.2387	0.0960	1.7918	0.5220
MEPS19	0.2403	0.0161	1.7511	0.1150
MEPS20	0.2485	0.0571	1.7936	0.3532
MEPS21	0.2528	0.0488	1.8686	0.3350
STAR	0.0230	0.0025	0.1605	0.0123
BIO	0.1051	0.0056	0.8509	0.0368
BLOG	0.3537	0.1135	2.3769	0.5421
BIKE	0.0921	0.0058	0.7759	0.0469

Figure 7: Length v.s. cubic metric. Larger cubic metric implies better efficiency (shorter band).

B.2 CERTIFYING CUBIC CONDITIONS

In this section, we validate the cubic conditions. The most important component for the cubic condition is Condition 2, which claims that conducting the quantile step would not hurt much efficiency. We next provide experiment results in Table 5 on comparing the average distance between each sample to their quantile in feature space, $M\|Q_1 - V^f_{D_{ca}}\| - \|V^f_{D_{ca}}\|$ and in output space $M[Q_1 - H(V^f_{D_{ca}}; D_{ca}) - H(V^f_{D_{ca}}; D_{ca})]$. We here take $= 1$ for simplicity. The significant gap in Table 5 validates that the distance in feature space is significantly smaller than that in output space, although we did not consider the Lipschitz factor L for computational simplicity.

Besides, we plot the relationship between the efficiency (band length) v.s. cubic metric in Figure 7. Specifically, cubic metric here represents the core statement in cubic condition (statement 2), which implies a metric form like $M\|Q_1 - V^f_{D_{ca}}\| - \|V^f_{D_{ca}}\|$. The results are shown in Figure 7.

B.3 FEATURE CP WORKS DUE TO SEMANTIC INFORMATION IN FEATURE SPACE

Experiment results illustrate that feature-level techniques improve the efficiency of conformal prediction methods (e.g., Feature CP vs. CP, Feature CQR vs. CQR). We claim that exploiting the semantic information in feature space is the key to our algorithm. Different from most existing conformal prediction algorithms, which regard the base model as a black-box mode, feature-level operations allow seeing the training process via the trained feature. This is novel and greatly broadens the scope of conformal prediction algorithms. For a well-trained base model, feature-level techniques improve efficiency by utilizing the powerful feature embedding abilities of well-trained neural networks.

In contrast, if the base model is untrained with random initialization (whose representation space does not have semantic meaning), Feature CP returns a similar band length as the baseline (see Table 6). This validates the hypothesis that Feature CP's success lies in leveraging the inductive bias of deep representation learning. Fortunately, realistic machine learning models usually contain meaningful information in the feature space, enabling Feature CP to perform well.

Table 6: Untrained base model comparison between conformal prediction and Feature CP. The base model is randomly initialized but not trained with the training fold. Experiment results show that Feature CP cannot outperform vanilla CP if the base model is not well-trained.

METHOD	VANILLA CP		FEATURE CP	
DATASET	COVERAGE	LENGTH	COVERAGE	LENGTH
COMMUNITY	90.28 1.70	4.85 0.22	90.68 1.33	4.92 0.77
FACEBOOK1	90.15 0.15	3.42 0.25	90.16 0.12	3.20 0.50
FACEBOOK2	90.17 0.11	3.51 0.26	90.12 0.14	3.34 0.39
MEPS19	90.81 0.46	4.02 0.16	90.86 0.30	4.22 0.48
MEPS20	90.10 0.60	4.10 0.28	90.28 0.46	4.02 0.41
MEPS21	89.78 0.44	4.08 0.16	89.85 0.58	3.81 0.32
STAR	90.07 0.77	2.23 0.18	89.47 1.84	2.24 0.40
BIO	90.06 0.19	4.25 0.11	90.11 0.07	4.44 0.74
BLOG	90.13 0.34	2.41 0.15	90.16 0.26	2.58 0.49
BIKE	89.53 0.78	4.65 0.15	89.61 0.86	4.13 0.38

Algorithm 4 Feature Conformalized Quantile Regression (Feature CQR)

Require: Level α, dataset $\mathcal{D} = \{(X_i, Y_i)\}_{i \in \mathcal{I}}$, test point X^0;

1: Randomly split the dataset $\mathcal{D}$ into a training fold $\mathcal{D}_{tr}$, $(X_i, Y_i)_{i \in \mathcal{I}_{tr}}$ together with a calibration fold $\mathcal{D}_{ca}$, $(X_i, Y_i)_{i \in \mathcal{I}_{ca}}$;

2: Train a base machine learning model $\hat{g}^{lo} \leftarrow \hat{f}^{lo}(\cdot)$ and $\hat{g}^{hi} \leftarrow \hat{f}^{hi}(\cdot)$ using $\mathcal{D}_{tr}$ to estimate the quantile of response Y_i, which returns $[\hat{Y}_i^{lo}, \hat{Y}_i^{hi}]$;

3: For each $i \in \mathcal{I}_{ca}$, calculate the index $c_i^{lo} = \mathbb{I}(\hat{Y}_i^{lo} \geq Y)$ and $c_i^{hi} = \mathbb{I}(\hat{Y}_i^{hi} \geq Y)$;

4: For each $i \in \mathcal{I}_{ca}$, calculate the non-conformity score $V_i^{lo} = \nabla_i^{lo} c_i^{lo}$ where ∇_i^{lo} is derived on the lower bound function with Algorithm 2;

5: Calculate the $(1-\alpha)$-th quantile Q_1^{lo} of the distribution $\frac{1}{|\mathcal{I}_{ca}|+1} \sum_{i \in \mathcal{I}_{ca}} V_i^{lo} + \infty$;

6: Apply Band Estimation on test data feature $\hat{u}(X^0)$ with perturbation Q_1^{lo} and prediction head $\hat{g}^{lo}$, which returns $[C_0^{lo}, C_1^{lo}]$;

7: Apply STEP 4-6 similarly with higher quantile, which returns $[C_0^{hi}, C_1^{hi}]$;

8: Derive $C_1^{fcqr}(X)$ based on Equation (12);

Ensure: $C_1^{fcqr}(X)$.

B.4 FEATURE CONFORMALIZED QUANTILE REGRESSION

In this section, we show that feature-level techniques are pretty general in that they can be applied to most of the existing conformal prediction algorithms. Specifically, We take Conformalized Quantile Regression (CQR, Romano et al. (2019)) as an example and propose Feature-level Conformalized Quantile Regression (Feature CQR). The core idea is similar to Feature CP (See Algorithm 3), where we conduct calibration steps in the feature space. We summarize the Feature CQR algorithm in Algorithm 4.

Similar to CQR, Algorithm 4 also considers the one-dimension case where $\mathcal{Y} \in \mathbb{R}$. We next discuss the steps in Algorithm 4. Firstly, different from Feature CP, Feature CQR follows the idea of CQR that the non-conformity score can be negative (see Step 4). Such negative scores help reduce the band length, which improves efficiency. This is achieved by the index calculated in Step 3. Generally, if the predicted value is larger than the true value $\hat{Y}_i^{lo} > Y_i$, we need to adjust $\hat{Y}_i^{lo}$ to be smaller, and vice visa. Step 8 follows the adjustment, where we summarize the criterion in Equation (12), given the two band $[C_0^{lo}, C_1^{lo}]$ and $[C_0^{hi}, C_1^{hi}]$.

[3] Here, we set $t = 1$ for simplicity

Table 7: Comparison between CQR and Feature CQR. Feature CQR achieves better ef ciency while maintaining effectiveness.

METHOD	CQR		FEATURE CQR	
DATASET	COVERAGE	LENGTH	COVERAGE	LENGTH
COMMUNITY	90:33 1:36	1:60 0:08	90:23 1:65	1.23 0:15
FACEBOOK1	89:94 0:23	1:15 0:04	92:00 0:20	1.00 0:04
FACEBOOK2	89:99 0:03	1:25 0:08	92:19 0:17	1.08 0:06
MEPS19	90:26 0:38	2:41 0:26	91:25 0:43	1.48 0:19
MEPS20	89:78 0:60	2:47 0:10	90:9 0:39	1.34 0:35
MEPS21	89:52 0:32	2:26 0:20	90:2 0:53	1.69 0:20
STAR	90:99 1:08	0:20 0:00	89:88 0:33	0.13 0:01
BIO	90:09 0:36	1:39 0:01	89:88 0:27	1.22 0:02
BLOG	90:15 0:15	1:47 0:05	91:49 0:26	0.89 0:03
BIKE	89:38 0:25	0:58 0:01	89:95 0:98	0.38 0:02

$$
\begin{aligned}
&\text{if } c_i^{lo} < 0, c_i^{hi} < 0, \text{ return } C_1^{fcqr}(X) = [C_0^{lo}, C_1^{hi}];\\
&\text{if } c_i^{lo} < 0, c_i^{hi} > 0, \text{ return } C_1^{fcqr}(X) = [C_0^{lo}, C_0^{hi}];\\
&\text{if } c_i^{lo} > 0, c_i^{hi} < 0, \text{ return } C_1^{fcqr}(X) = [C_1^{lo}, C_1^{hi}];\\
&\text{if } c_i^{lo} > 0, c_i^{hi} > 0, \text{ return } C_1^{fcqr}(X) = [C_1^{lo}, C_0^{hi}]:
\end{aligned}
\tag{12}
$$

Similar to Feature CP, we need a Band Estimation step to approximate the band length used in Step 6. One can change it into Band Detection if necessary. Different from Feature CP where Band Estimation always returns the upper bound of the band, Feature CQR can only approximate it. We conduct experiments to show that this approximation does not lose effectiveness since the coverage is always approximate to $1 \quad$. Besides, different from CQR, which considers adjusting the upper and lower with the same value, we adjust them separately, which is more exible in practice (see Step 7).

We summarize the experiments result in Table 7. Feature CQR achieves better ef ciency while maintaining effectiveness. Here we provide 90% con dence band using ve repeated experiments with different random seeds.

B.5 GROUP COVERAGE FOR FEATURE CONFORMALIZED QUANTILE REGRESSION

This section introduces the group coverage returned by feature-level techniques, which implies the performance conditional coverage, namely $\mathbb{P}(Y \in C(X) | X)$. Speci cally, we split the test set into three groups according to their response values, and report the minimum coverage over each group.

We remark that the group coverage of feature-level conformal prediction stems from its vanilla version. That is to say, when the vanilla version has a satisfying group coverage, its feature-level version may also return a relatively satisfying group coverage. Therefore, we did not provide Feature CP here because vanilla CP cannot return a good group coverage.

We summarize the experiment results in Table 8. Although we did not provide a theoretical guarantee for group coverage, Feature CQR still outperforms vanilla CQR in various datasets in terms of group coverage. Among ten datasets, Feature CQR outperforms vanilla CQR in four datasets, and is comparable with vanilla CQR in ve datasets. Although the advantage is not universal, improving group coverage via feature-level techniques is still possible.

We note that there is still one dataset where vanilla CQR outperforms Feature CQR. We attribute the possible failure reason of Feature CQR on the dataset FACEBOOK2 to the failure of base models. As stated in Section B.3, Feature CQR only works when the base model is well-trained. However, when grouping according to the returned values, it is possible that there exists one group that is not well-trained during the training process. This may cause the failure of Feature CQR on the dataset FACEBOOK2.

Table 8: Comparison between Feature CQR and CQR in terms of group coverage.

GROUP	COVERAGE	FEATURE CQR		VANILLA CQR	
COMMUNITY		76.05	5.13	78.77	4.17
FACEBOOK1		65.52	0.95	66.68	1.69
FACEBOOK2		65.66	1.41	70.78	1.39
MEPS19		76.67	2.17	71.26	1.23
MEPS20		77.70	0.90	71.26	3.20
MEPS21		74.71	2.36	70.74	1.83
STAR		84.62	2.77	82.20	5.72
BIO		84.80	1.05	80.03	1.48
BLOG		59.43	0.60	49.10	0.54
BIKE		81.07	1.65	78.22	2.44

Table 9: Ablation study of the number of layers f and g ($\alpha = 0.1$) in unidimensional tasks, where the default setting is $f : 2, g : 2$.

DATASET	BIO				BIKE				BLOG			
METHOD	COVERAGE		LENGTH		COVERAGE		LENGTH		COVERAGE		LENGTH	
$f : 2$ $g : 2$	90.20	0.39	1.873	0.06	89.61	0.94	1.794	0.11	90.06	0.44	2.811	0.54
$f : 3$ $g : 1$	90.24	0.32	1.961	0.02	89.72	1.10	1.917	0.14	90.16	0.34	3.319	0.22
$f : 1$ $g : 3$	90.00	0.46	1.860	0.12	89.72	0.81	1.748	0.10	90.11	0.43	2.595	0.38

B.6 ROBUSTNESS OF SPLITTING POINT

As discussed in the main text, the empirical coverage of Feature CP is pretty robust to the splitting point. We show both experiments of small neural networks (Table 9) and large neural networks (Table 10). We remark that one can also apply standard cross-validation to find the best splitting point, which may return a shorter band.

B.7 FCP IN CLASSIFICATION

In this section, we show how to deploy FCP in classification problems. The basic ideas follow Algorithm 3, except for two points:

1. We use cross-entropy loss instead of MSE loss when calculating the non-conformity score.
2. We do not deploy LiPRA, but use a sampling method to return the final confidence band, which induces a looser bound.

For each sample in the calibration fold, we sample 1000 samples in the confidence band of feature space, and return the corresponding predictions in output space to get the final confidence band. We summarize the experiment results in Table 2, where we follow the statistics of APS and RAPS in Angelopoulos et al. (2021b).

B.8 THE VALIDATION OF NON-CONFORMITY SCORE

Distribution for non-conformity score in calibration fold. We plot the distribution of calibration score in Figure 8. We plot each non-conformity score in the calibration fold. The distribution of non-conformity scores is smooth and single-peak in the real-world dataset, meaning that the proposed score is reasonable.

Besides, one may wonder what happens if the predicted function is close to the ground truth function, which may lead to a zero quantile of the non-conformity score. We conduct experiments to show that zero quantiles would not underestimate the final non-conformity score. Specifically, we create a dataset that the prediction target is given by $y = f(x) + \epsilon$, where f denotes a three-layer neural network, $\epsilon = 0$ with probability 0.901 and ϵ follows a standard Gaussian with probability 0.099.

Table 10: Ablation study of the number of layers in g ($\alpha = 0.1$) in large neural networks on ImageNet.

SPLITTING POINT (PREDICTION HEAD g)	COVERAGE		LENGTH	
THE LAST LAYER	0.900	0.0018	3.32	0.10
LAST TWO LAYERS	0.901	0.0047	3.26	0.03
LAST THREE LAYERS	0.900	0.0029	3.30	0.01

Figure 8: The distribution of the calibration score for the segmentation task. This indicates that the definition of our non-conformity score is a proper one.

Here the 0.901 is to let the 90% quantile of the non-conformity score on a calibration set to be 0. Besides, we make the predicted model satisfy $\hat{f} = f$. We summarize the results (with $\alpha = 0.1$) as follows (five repeated experiments with different $\mathbf{x}$). Here $\mathbf{x}$ is generated with a uniform distribution between 0 and 1. Note that the zero-length is implied by zero quantiles.

However, in practice it is impossible for the predicted model to be f, therefore exact zero-quantile is impossible. We then summarize the results when the model $\hat{f}$ is obtained via training on a training fold. The non-zero length implies that the quantile is non-zero.

B.9 ADDITIONAL EXPERIMENT RESULTS

This section provides more experiment results omitted in the main text.

Visualization for the segmentation problem. We also provide more visualization results for the segmentation problem in Figure 9.

Table 11: Zero quantile does not effects the empirical coverage of $\hat{r}_{\text{avg}}(\cdot)$.

TRIAL	COVERAGE	LENGTH
1	90.63	0
2	90.08	0
3	89.72	0
4	89.66	0
5	90.21	0

Table 12: Zero quantile does not effects the empirical coverage of $\hat{r}_{\text{avg}}(\cdot)$.

TRIAL	COVERAGE	LENGTH
1	90.55	0.1274
2	89.72	0.1105
3	89.75	0.1120
4	89.31	0.1202
5	90.10	0.1245

Figure 9: More visualization results for the Cityscapes segmentation task.