CP^2: Leveraging Geometry for Conformal Prediction via Canonicalization

Putri A. van der Linden[*1] **Alexander Timans**[1,2] **Erik J. Bekkers**[1]

[1]Amsterdam Machine Learning Lab, University of Amsterdam
[2]UvA-Bosch Delta Lab, University of Amsterdam

Abstract

We study the problem of *conformal prediction* (CP) under geometric data shifts, where data samples are susceptible to transformations such as rotations or flips. While CP endows prediction models with *post-hoc* uncertainty quantification and formal coverage guarantees, their practicality breaks under distribution shifts that deteriorate model performance. To address this issue, we propose integrating geometric information—such as geometric pose—into the conformal procedure to reinstate its guarantees and ensure robustness under geometric shifts. In particular, we explore recent advancements on pose *canonicalization* as a suitable information extractor for this purpose. Evaluating the combined approach across discrete and continuous shifts and against equivariant and augmentation-based baselines, we find that integrating geometric information with CP yields a principled way to address geometric shifts while maintaining broad applicability to black-box predictors.

1 INTRODUCTION

The deployment of machine learning models—including deep neural networks—has become increasingly widespread, yet their application in costly or safety-critical settings remains hindered by two key challenges [Makridakis and Bakas, 2016, Quiñonero-Candela et al., 2022]. Firstly, many models continue to produce point-wise predictions without uncertainty estimation, inherently limiting the robustness of obtained information for decision-making [Begoli et al., 2019, Padilla et al., 2021]. Yet even when uncertainty is incorporated, such as through some form of uncertainty scoring or probabilistic modelling [Gawlikowski et al., 2023], estimates can be misleading or overconfident [Kompa et al., 2021, Xiong et al., 2023]. A popularized uncertainty framework that partially addresses such issues is *conformal prediction* (CP), which extends point-wise predictions to prediction set or interval estimation [Vovk et al., 2005, Angelopoulos et al., 2023]. Importantly, a notion of reliability is obtained via a probabilistic coverage guarantee for new, unseen test samples (see § 2.1). Unlike traditional prediction set methods [Khosravi et al., 2011], CP is fully data-driven, distribution-free, and compatible with 'black-box' models.

Secondly, it is well-known that distribution shifts at test time can severely degrade model performance [Koh et al., 2021, Ovadia et al., 2019]. Among types of shifts, *geometric* data shifts—where test samples undergo geometric transformations such as rotations or flips—pose a significant challenge, in particular for pretrained models lacking integrated equivariance or invariance properties [Bronstein et al., 2021]. As such symmetry-awareness can be sometimes challenging to scale and is thus overlooked [Brehmer et al., 2024], large models trained on vast datasets may nonetheless struggle when faced with pose variations, as exemplified in Tab. 1 for segmentation under rotations. Other practical failures may include proper recognition for medical images due to scan variations [Fu et al., 2023] or 3D objects due to axis-misaligned point clouds [Vadgama et al., 2025]. For conformal prediction, such geometric shifts can violate *exchangeability* assumptions on the data (Def. 3.1), leading to potentially unreliable or uninformative prediction sets [Barber et al., 2023]. Unreliable in the sense that statistical coverage guarantees may no longer hold, and uninformative as prediction sets may grow excessively large.

To address this, we propose robustifying the conformal procedure by incorporating geometric information on occuring shifts, while preserving CP's advantageous flexibility by avoiding to modify the underlying model. This is practically achieved via *canonicalization* [Mondal et al., 2023, Kaba et al., 2023], a framework that learns to map data into a canonical form, and decouples the geometric task from the underlying predictor. Leveraging this approach, we explore

[*]Corresponding author: p.a.vanderlinden@uva.nl

Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI 2025), PMLR 244:2642–2658.

Table 1: Zero-shot Mask-RCNN segmentation performance (mAP) on regular and $C4$-rotated COCO data without and with invariance (via canonicalization [Mondal et al., 2023]). Missing symmetry-awareness leads to failed generalization.

Model	mAP	$C4$-mAP
Mask-RCNN without Invariance	47.81	12.79
Mask-RCNN with Invariance	43.47	43.47

how obtained geometric information can be effectively combined with CP in multiple different ways. In summary, our contributions include:

- Introducing a novel geometric perspective on the topic of distribution shifts in conformal prediction, and motivating how geometric information can ensure core conditions of CP such as exchangeability are met (§ 3);
- Leveraging canonicalization as a suitable geometric information extractor that is both *post-hoc* and lightweight, in line with practical principles underlying CP;
- Investigating its integration with CP in several ways, including mitigating performance drops (§ 4.1), as an information tool for conditional coverage (§ 4.2), and as a weighting mechanism in multi-shift settings (§ 4.3).

2 BACKGROUND

We next provide some background on conformal prediction (§ 2.1), group equivariance and invariance properties (§ 2.2), and the canonicalization framework (§ 2.3). Regarding notation, let $\mathcal{X} \times \mathcal{Y}$ mark the sample space with some data-generating distribution P over it, and $\mathbf{x}, \mathbf{y}$ random variables with realizations $\boldsymbol{x}, \boldsymbol{y}$. We denote any learnable functions, such as a prediction model $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$, as mappings with learnable parameters $\boldsymbol{\theta} \in \Theta$.

2.1 CONFORMAL PREDICTION

We consider the usual setting of *split conformal prediction*[1], wherein a hold-out calibration set $\mathcal{D}_{cal} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ and test set $\mathcal{D}_{test} = \{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}_{j=n+1}^{n+m}$ are both sampled exchangeably (*i.e.* permutation invariantly, see Def. 3.1) from some fixed distribution P_0 [Papadopoulos et al., 2007]. Using a pre-specified scoring function $s : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and pretrained predictor $f_{\boldsymbol{\theta}}$, we compute a set of nonconformity scores $S = \{s_i\}_{i=1}^{n}$ on $\mathcal{D}_{cal}$, where $s_i = s(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \boldsymbol{y}_i)$. These scores encode a desired notion of disagreement between predictions and responses, such as a simple residual score $s_i = |f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \boldsymbol{y}_i|$ for regression or predicted probability $s_i = 1 - p(\mathbf{y}_i = \boldsymbol{y}_i|\boldsymbol{x}_i)$ for classification. Next, a sample-corrected conformal quantile $Q_{1-\alpha}(F_S)$ is computed, where F_S denotes the empirical distribution over the calibration scores[2], and $\alpha \in (0, 1)$ a tolerated miscoverage rate. Given a new test sample $(\boldsymbol{x}_{n+1}, \boldsymbol{y}_{n+1})$, a prediction set is then constructed as $C(\boldsymbol{x}_{n+1}) = \{\boldsymbol{y} \in \mathcal{Y} : s(f_{\boldsymbol{\theta}}(\boldsymbol{x}_{n+1}), \boldsymbol{y}) \leq Q_{1-\alpha}(F_S)\}$, *i.e.*, we include candidate responses whose score does not exceed the quantile. Exploiting the data's exchangeability under P_0, a formal coverage guarantee on inclusion of the true response $\boldsymbol{y}_{n+1}$ can then be given with high probability as

$$\mathbb{P}(\boldsymbol{y}_{n+1} \in C(\boldsymbol{x}_{n+1})) \geq 1 - \alpha. \tag{1}$$

We refer to Shafer and Vovk [2008], Angelopoulos et al. [2024b] for details on the intuition and technical proofs.

Mondrian conformal prediction. The coverage guarantee in Eq. 1 only holds *marginally* over $\mathcal{D}_{cal} \cup \mathcal{D}_{test}$, thus ensuring coverage in a broad sense. Stronger and more refined guarantees can be obtained by simply partitioning the data into sub-populations of interest, and running the conformal procedure per partition. We refer to this as *partition-conditional* or *mondrian* conformal prediction [Toccaceli and Gammerman, 2019]. If we consider a mapping $\phi : \mathcal{X} \times \mathcal{Y} \to \{1, \ldots, K\}$ assigning each sample to a data partition, the coverage guarantees hold per partition as

$$\mathbb{P}(\boldsymbol{y}_{n+1} \in C(\boldsymbol{x}_{n+1}) \,|\, \phi(\boldsymbol{x}_{n+1}, \boldsymbol{y}_{n+1}) = k) \geq 1 - \alpha \tag{2}$$

for all $k \in \{1, \ldots, K\}$. Data partitions of interest can include distinction by class label [Cauchois et al., 2021], feature properties [Sesia and Romano, 2021a, Jung et al., 2023], or a balancing criterion like fairness [Romano et al., 2020a].

Weighted conformal prediction. To enhance data adaptivity and address settings of reduced exchangeability, a weighted formulation of CP is given by replacing the conformal quantile with $Q_{1-\alpha}(\tilde{F}_S)$, where $\tilde{F}_S$ now denotes the empirical distribution over a *weighted* score set as

$$\tilde{F}_S = \sum_{i=1}^{n} \tilde{w}_i \cdot \delta(s_i) + \tilde{w}_{n+1} \cdot \delta(+\infty), \tag{3}$$

with $\delta(s_i)$ denoting the dirac delta centered at score s_i, and $\tilde{w}_i$ its associated normalized weight such that $\sum_{i=1}^{n} \tilde{w}_i = 1$. For example, Barber et al. [2023] suggest fixed weighting schemes such as upweighting more recent samples in a data stream setting, while Guan [2023] propose data-dependent (unnormalized) weights guided by feature distances such as the kernel distance $w_i = \exp\{-h\,|\boldsymbol{x}_i - \boldsymbol{x}_{n+1}|\}$.

2.2 GROUP EQUIVARIANCE AND INVARIANCE

Formally, we denote a symmetry group G as a set of elements with a binary operator $\cdot$ satisfying closure and associativity, and for which an identity element e and inverses

[1] As opposed to full or cross-validation conformal schemes.

[2] Extended with $\{+\infty\}$ to ensure proper coverage adjustments.

g^{-1} exist such that $e \cdot g = g$ and $g^{-1} \cdot g = e$ respectively [Cohen and Welling, 2016]. In our context, G can be described as a structured space of possible symmetry transformations on the data. That is, a sample $\mathbf{x} \in \mathcal{X}$ is transformed by a *group action* as $\rho(g) \cdot \mathbf{x}$, where $g \in G$ denotes a group element and $\rho : G \to T$ a group representation mapping g to a concrete transformation[3]. For instance, if we define $G = SO(2)$ as the group of planar rotations, then g might represent a particular rotation angle, and $\rho(g)$ the rotation of $\mathbf{x}$ by that angle via matrix multiplication. Given such geometric data transformations, desirable properties for some predictor $f_{\boldsymbol{\theta}}$ can include *(i)* preserving the symmetry structure of G by commuting with group actions, *i.e.* being *equivariant*, or *(ii)* ensuring robustness to group actions by remaining *invariant* to them. Specifically, $f_{\boldsymbol{\theta}}$ is deemed group equivariant if for all $g \in G$ we have that

$$f_{\boldsymbol{\theta}}(\rho(g) \cdot \mathbf{x}) = \rho'(g) \cdot f_{\boldsymbol{\theta}}(\mathbf{x}), \tag{4}$$

where $\rho(g)$ and $\rho'(g)$ act on the data input space $\mathcal{X}$ and output space $\mathcal{Y}$, respectively. Thus the model's output commutes predictably with the applied transformation, a property frequently employed, for instance, in translation-equivariant convolutional models for image processing. In contrast, if $\rho'(g) = \mathbb{I}$ equates the identity transformation for any group element g, then $f_{\boldsymbol{\theta}}$ is group-invariant to G. This property is desirable if input samples $\mathbf{x}$ are subject to geometric data transformations or shifts, but we desire $f_{\boldsymbol{\theta}}$ to provide consistent prediction outputs regardless. In neural network models, both properties are typically achieved by employing architectures that inherently incorporate Eq. 4 as a constraint, or through explicit or implicit learning of symmetries, *e.g.* via data augmentation (see § 5).

2.3 EQUIVARIANCE VIA CANONICALIZATION

Instead of designing a model and its layers to be equivariant, one may also obtain equivariance through *canonicalization* [Mondal et al., 2023, Kaba et al., 2023]. At its core, canonicalization aims to learn a mapping from potentially transformed data to its standardized or canonical orientation before processing by the predictor. The approach separates the tasks of correcting and predicting for transformed data, greatly increasing flexibility by allowing the use of *non*-equivariant pretrained predictors within an equivariant framework. More formally, given a predictor $f_{\boldsymbol{\theta}}$ we additionally consider a learnable *canonicalization network* (CN) as $c_{\boldsymbol{\theta}} : \mathcal{X} \to G$, and denote the canonicalization process as

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = \rho'(c_{\boldsymbol{\theta}}(\mathbf{x})) \cdot f_{\boldsymbol{\theta}}\left(\rho(c_{\boldsymbol{\theta}}(\mathbf{x})^{-1}) \cdot \mathbf{x}\right). \tag{5}$$

The CN $c_{\boldsymbol{\theta}}$ aims to predict the (inverse) group element to map $\mathbf{x}$ back to its canonical form, and Eq. 5 ensures $f_{\boldsymbol{\theta}}$ is $G-$equivariant if $c_{\boldsymbol{\theta}}$ itself is $G-$equivariant [Kaba et al., 2023]. Similarly, for invariance we have $\rho'(g) = \mathbb{I}$ and Eq. 5 simplifies to

$$f_{\boldsymbol{\theta}}(\mathbf{x}) = f_{\boldsymbol{\theta}}\left(\rho(c_{\boldsymbol{\theta}}(\mathbf{x})^{-1}) \cdot \mathbf{x}\right) = f_{\boldsymbol{\theta}}\left(\hat{g}^{-1} \cdot \mathbf{x}\right), \tag{6}$$

where we've omitted ρ since there is no ambiguity on the group action space[4], and $\hat{g}$ denotes the predicted group element using $c_{\boldsymbol{\theta}}$. Whereas the original formulation by Kaba et al. [2023] directly predicts a single group element $\hat{g} = c_{\boldsymbol{\theta}}(\mathbf{x})$, Mondal et al. [2023] extend the approach to predict a group distribution $\hat{P}_{G|\mathbf{x}}$ over transformations, in which case $\hat{g} \sim \hat{P}_{G|\mathbf{x}}$ can be sampled.

Regularization using the canonicalization prior. There are practical challenges in ensuring that the learning process of the CN is both coupled to the employed predictor and the correct poses in the data. Thus, Mondal et al. [2023] propose training the CN with a double objective of the form $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \beta \cdot \mathcal{L}_{\text{prior}}$, where $\mathcal{L}_{\text{task}}$ is a cross-entropy loss term and $\mathcal{L}_{\text{prior}}$ a regularization term. In particular, if $f_{\boldsymbol{\theta}}$ is pretrained and fully frozen during training the task loss is zero, and an additional learning signal becomes necessary. Thus, the *canonicalization prior* (CP) term $\mathcal{L}_{\text{prior}}$ is introduced to align the CN's learned poses with the canonical pose prevalent in the data $\mathcal{D}_{can} \sim P_{can}$ used to learn the CN (*e.g.* a hold-out data split). The loss is then given by

$$\mathcal{L}_{\text{prior}} = \mathbb{E}_{P_{can}}[D_{KL}(P_{G|\mathbf{x}} \,||\, \hat{P}_{G|\mathbf{x}})], \tag{7}$$

where $P_{G|\mathbf{x}}$ is a prior distribution for the group elements acting on samples in $\mathcal{D}_{can}$, and D_{KL} the Kullback-Leibler divergence. In practice the prior is usually set to $P_{G|x} = \delta(e)$, *i.e.*, full probability mass on the identity element, thus assuming the 'correct' data is subject to no transformations. This additionally simplifies computation of Eq. 7 for particular groups, *e.g.* for discrete rotations we obtain $\mathcal{L}_{\text{prior}} = -\mathbb{E}_{P_{can}} \log \hat{P}_{G|\mathbf{x}}(e)$, the negative log probability of the identity element [Mondal et al., 2023]. Note that G still needs to be defined beforehand, *i.e.* the CN learns a distribution *over* group elements, rather than a set of valid group elements themselves (from a possibly infinite space). However, we find that results are not overly impacted when the correct group is a subgroup $G' \subset G$ of the model-specified group (*e.g.*, $C4$ rotations rather than $C8$ rotations), providing some leeway to misspecification (see Tab. 3).

3 GEOMETRIC INFORMATION FOR CONFORMAL PREDICTION

We next motivate why canonicalization suits itself naturally for joint use with conformal prediction, including a perspective on data exchangeability. In § 3.1 we then outline three ways to leverage obtained geometric information for conformal procedures under differing shift scenarios.

[3] $T \subset GL(V)$ denotes a subset of the total set of linear invertible transformations on some vector space V.

[4] And subsequently abuse notation for simplicity and use $g \cdot \mathbf{x}$ as the application of g on the domain directly.

Practical motivation: flexible and efficient. Equivariance modelling usually requires custom prediction models which embed the necessary geometric constraints deep within their architecture, such as via group convolutions with regular [Cohen and Welling, 2016, Bekkers, 2020] or steerable filters [Weiler and Cesa, 2019]. This introduces additional complexity into the model, complicates training, and can hamper the transferability of a solution across datasets or tasks. In contrast, canonicalization effectively decouples the prediction and equivariance components, permitting the use of a broader variety of non-equivariant, pretrained models for prediction, and ensuring equivariance in a *post-hoc* step. This outsourcing permits the use of more efficient, light-weight equivariant models to learn the canonical mapping in an unsupervised way, while the complex prediction task is handled by a separate, usually substantially larger model (magnitudes larger, see § 4.1). This can also provide benefits over data augmentation, since only a single forward pass through the predictor is necessary. Most crucially, the obtained flexibility meshes particularly well with the conformal prediction framework, as CP's key advantage of *post-hoc* compatibility with arbitrary 'black-box' predictors is preserved. In that sense, we may think of canonicalization as a second 'bolt-on' module, situated inbetween the predictor and uncertainty estimation via CP. Naturally, canonicalization has little to no effect on models that are *already* symmetry-aware, as the additional module then becomes redundant.

Theoretical motivation: canonical mapping as data exchangeability. We may also motivate canonicalization for CP from a more fundamental data perspective. Intuitively, the canonicalization network $c_{\boldsymbol{\theta}}$ aids mitigate the predictor's performance loss due to encountered geometric shifts by enforcing data exchangeability with the training set, in turn benefitting uncertainty estimation. More formally, let us first define *data exchangeability* following the CP framework:

Definition 3.1 (Exchangeability [Shafer and Vovk, 2008])**.** *A sequence of random variables $\mathbf{x}_1, \dots, \mathbf{x}_n$ is exchangeable if for any permutation $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$ with $n \geq 1$ we have that $P(\mathbf{x}_1, \dots, \mathbf{x}_n) = P(\mathbf{x}_{\pi(1)}, \dots, \mathbf{x}_{\pi(n)})$.*

That is, the joint data probability is invariant to sample ordering. In particular, observe how the *i.i.d* setting is a special case where $P(\mathbf{x}_1, \dots, \mathbf{x}_n)$ factorizes. For conformal coverage guarantees along Eq. 1 to nominally hold, Def. 3.1 only needs to be satisfied across calibration ($\mathcal{D}_{cal}$) and test data ($\mathcal{D}_{test}$), but *not* necessarily for the predictor's training set ($\mathcal{D}_{train}$). However, learned data properties that poorly translate to new (shifted) samples will result in a low-quality set of computed nonconformity scores, starkly inflating prediction set sizes and rendering obtained sets uninformative. From a data perspective, this issue can be alleviated if $\mathcal{D}_{train}$ also approximately satisfies Def. 3.1, and $f_{\boldsymbol{\theta}}$ thus guarantees informative scoring.

This is precisely what the CN attempts to ensure via its canonical mapping. Classical exchangeability imposes data invariance under permutations $\pi \in \mathbb{S}_n$, where $\mathbb{S}_n$ represents the set of all permutations in $\{1, \dots, n\}$. Assuming a shift by the group G affecting $\mathcal{D}_{cal}$, each calibration sample $\mathbf{x}_i$ is now also susceptible to an independent transformation $g_i \in G$. That is, on a dataset level we now aim for exchangability (*i.e.* group invariance) to extend to the group $G^n = G \times G \times \dots \times G$, in which each sample experiences a potentially different transformation of G. For every affected sample $g_i \cdot \mathbf{x}_i$, the CN ensures the existence of an inverse transform $c_{\boldsymbol{\theta}}(g_i \cdot \mathbf{x}_i)^{-1}$ which neutralizes g_i. That is, proper canonicalization maintains the relationship $c_{\boldsymbol{\theta}}(g \cdot \mathbf{x})^{-1} = c_{\boldsymbol{\theta}}(\mathbf{x})^{-1} \cdot g^{-1}$ for all $g \in G$ and inputs $\mathbf{x}$ [Kaba et al., 2023]. Under the action of G^n, we then observe for the joint distribution that

$$\begin{aligned} P(c_{\boldsymbol{\theta}}(g_1 \cdot \mathbf{x}_1)^{-1} \cdot g_1 \cdot \mathbf{x}_1, \dots, c_{\boldsymbol{\theta}}(g_n \cdot \mathbf{x}_n)^{-1} \cdot g_n \cdot \mathbf{x}_n) \\ = P(c_{\boldsymbol{\theta}}(\mathbf{x}_1)^{-1} \cdot \mathbf{x}_1, \dots, c_{\boldsymbol{\theta}}(\mathbf{x}_n)^{-1} \cdot \mathbf{x}_n)\,, \end{aligned}$$

ensuring that the distribution over canonicalized samples remains invariant under G^n. This generalizes the classical exchangeability definition of Def. 3.1 to include both dataset permutations and sample-wise transformations, enlarging the symmetry group from $\mathbb{S}_n$ to $\mathbb{S}_n \times G^n$. Since distributional invariance to G implies that transformations from G do not alter the joint distribution, the CN effectively enforces *probabilistic symmetry* (Bloem-Reddy et al. [2020], Prop. 1). Thus, it guarantees well-calibrated nonconformity scores practically useful for CP even under geometric shifts.

3.1 USE CASES FOR CONFORMAL PREDICTION

Following our motivation, we now illustrate three interesting ways how obtained group information can be leveraged to benefit different conformal prediction procedures and tasks.

For general robustness to geometric data shifts. We first directly demonstrate the obtained robustness to a geometric data shift at calibration and test time. To that end, we can simply combine the CN $c_{\boldsymbol{\theta}}$ with a non-equivariant, pretrained predictor and apply standard split conformal prediction (SCP). Since the CN ensures the necessary exchangeable mapping to align the predictor's outputs with the conformal procedure, we expect a substantial improvement in prediction set sizes over directly using $f_{\boldsymbol{\theta}}$ and SCP without canonicalization.

As a diagnostics tool and proxy for conditional coverage. Unlike inherently equivariant models or models trained with data augmentation, canonicalization provides us with explicit access to *sample-wise* geometric information or pose via the group distributions $\hat{P}_{G|\mathbf{x}}$. These can be exploited to construct empirical group distributions pertaining to any separable data partition of interest, *e.g.* by class labels or feature

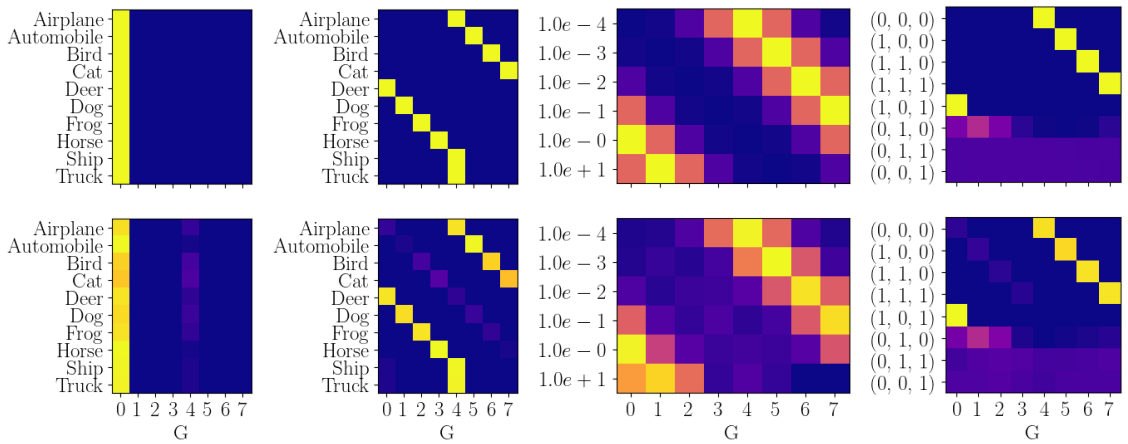


Figure 1: True (*top*) and recovered (*bottom*) partition-conditional group distributions based on different data partitions (class, entropy, color) and shifts (outlined in § A.2). *From left to right:* (a) Class partition, no shift; (b) Class partition, `dirac` shift; (c) Entropy partition, `normal` shift; (d) Color partition, `var-gauss` shift. For (c) samples are partitioned by predictive entropy into exp. scaled bins, for (d) samples are assigned based on RGB values and average image color. Both are used as a measure of sample complexity. We consider G as the $C8$ rotation group. Using the canonicalization network's group distributions and Eq. 8, the built group maps (*bottom*) accurately uncover existing partition-conditional geometric patterns.

properties. Such empirical group distributions can provide insights into the geometric poses under which a certain property or partition naturally occurs in the data (see § A.2 for further intution). If a partition's 'group map'—visualized for some examined partitions in Fig. 1—reveals informative geometric patterns, the group assignments can be subsequently leveraged to provide stronger partition-conditional or mondrian conformal prediction (MCP) guarantees (Eq. 2). In that sense, the group information can be leveraged as a data diagnostics tool to uncover even *a priori* unknown but geometrically informative partitions, or to suggest data exchangeability for a partition when no meaningful group pattern emerges. In principle, such group maps could be extended as far as incorporating multiple datasets to potentially uncover geometric shifts across new data sources.

Formally, given some data partition $k \in \{1, \dots, K\}$ of $\mathcal{D}_{cal}$ into K parts, an empirical group distribution for the k-th partition can be constructed as $\hat{P}_{G|k} = \{\hat{P}_{g|k} \mid g \in G\}$, where $\hat{P}_{g|k}$ denotes the g-th element's estimated frequency computed as

$$\hat{P}_{g|k} = \frac{\sum_{i=1}^{n} \mathbb{1}(\hat{g}_i = g \wedge \phi(\boldsymbol{x}_i, \boldsymbol{y}_i) = k)}{\sum_{i=1}^{n} \mathbb{1}(\phi(\boldsymbol{x}_i, \boldsymbol{y}_i) = k)}. \tag{8}$$

The indicator function is given by $\mathbb{1}[\cdot]$, whereas $\hat{g}_i \sim \hat{P}_{G|\boldsymbol{x}_i}$ is the sampled group element obtained for $(\boldsymbol{x}_i, \boldsymbol{y}_i)$.

As a weighting scheme for double shift settings. Consider a more complex *double shift* setting, wherein the first shift between $\mathcal{D}_{train}$ and $\mathcal{D}_{cal}$ is addressed by the CN, but an additional second shift between $\mathcal{D}_{cal}$ and $\mathcal{D}_{test}$ occurs. For example, the CN trained on $\mathcal{D}_{cal}$ learns to address a shift caused by the $C8$ rotation group, but test samples are susceptible to continuous rotations on $SO(2)$. In this case even the use of canonicalization with standard SCP can be insufficient to ensure conformal guarantees, since the CN can underperform when faced with new, unknown group elements (*i.e.* rotation angles in $SO(2)$ but not $C8$). However, the obtained group information can still be leveraged to inform *geometric weights* for a weighted conformal prediction strategy (WCP). We posit that the CN assigns higher probability to group elements that are 'closer' aligned with the test sample's unknown transformation, and as such provides information to upweigh more geometrically relevant calibration samples; we elaborate on this intuition in § A.3. In conjunction with WCP, this may offer improved robustness against shifts with *unknown* group elements. Different double-shift settings and approaches to establish robustness are outlined in Tab. 2[5].

Table 2: Different possible geometric shift settings for calibration and test data. Note that the CN is trained on $\mathcal{D}_{cal}$ and thus learns mappings to G. The density ratio $w(\mathbf{x})$ denotes reweighting on the same group support, while G_{new} denotes a group with newly encountered group elements, *i.e.* $G \subset G_{new}$. The train data for the predictor $f_{\boldsymbol{\theta}}$ in all cases is unaffected by geometric shift, *i.e.* $G_{train} = \{e\}$.

Train	Calibration	Test	Robustness
$\delta(e)$	$P_{G\|\mathbf{x}}$	$P_{G\|\mathbf{x}}$	CN + SCP
$\delta(e)$	$P_{G\|\mathbf{x}}$	$w(\mathbf{x}) \cdot P_{G\|\mathbf{x}}$	CN + SCP
$\delta(e)$	$P_{G\|\mathbf{x}}$	$P_{G_{new}\|\mathbf{x}}$	CN + WCP

Formally, given a test instance $\boldsymbol{x}_{n+1}$, the i-th calibration sample's geometric relevance with respect to $\boldsymbol{x}_{n+1}$ can be measured by $D\big(\hat{P}_{G|\boldsymbol{x}_{n+1}}, \hat{P}_{G|\boldsymbol{x}_i}\big)$, with D being any distri-

[5]We empirically examine the first (§ 4.1) and last rows (§ 4.3).

Table 3: Results with APS on CIFAR-100 for target coverage $(1-\alpha) = 95\%$ across different rotation shifts. Target coverage is efficiently maintained when no shift occurs, but grows excessively large and uninformative ( ) when the underlying predictor $\hat{f}_{\boldsymbol{\theta}}$ is not equivariant, or if the wrong group is specified (learning $C4$ but exposed to $C8$). Results are reported across $T = 10$ random calibration/test splits. # Param. indicates the required number of training parameters. CP^2 employs the same (pretrained and frozen) predictor $\hat{f}_\theta$, and only requires training a substantially smaller canonicalization network.

		No Shift			$C4$ **Rotation Shift**			$C8$ **Rotation Shift**		
Model	# Param.	Acc	Coverage	Set Size	Acc	Coverage	Set Size	Acc	Coverage	Set Size
$\hat{f}_\theta$	23.7 M	71.66	$95.09 \pm .003$	$6.212 \pm .106$	40.17	$96.49 \pm .002$	$55.149 \pm .388$	33.68	$95.63 \pm .001$	$61.402 \pm .524$
$\hat{f}_\theta$ + $SO(2)$ Aug.	23.7 M	60.13	$95.02 \pm .005$	$11.362 \pm .393$	59.72	$95.20 \pm .003$	$11.634 \pm .302$	58.27	$95.00 \pm .004$	$11.213 \pm .320$
$\hat{f}_\theta$ + C_4 Aug.	23.7 M	63.03	$95.05 \pm .006$	$10.047 \pm .475$	62.72	$95.11 \pm .005$	$10.175 \pm .452$	49.72	$98.10 \pm .001$	$88.892 \pm .249$
$\hat{f}_\theta$ + C_8 Aug.	23.7 M	62.53	$95.38 \pm .005$	$10.758 \pm .361$	62.37	$95.33 \pm .002$	$10.396 \pm .183$	60.82	$95.26 \pm .002$	$10.970 \pm .156$
CP^2 ($G = 4$)	1.0 M	65.37	$95.05 \pm .004$	$10.583 \pm .431$	65.37	$95.02 \pm .004$	$10.557 \pm .443$	48.19	$95.02 \pm .005$	32.605 ± 2.382
CP^2 ($G = 8$)	2.0 M	65.46	$95.34 \pm .004$	$11.198 \pm .398$	65.46	$94.95 \pm .004$	$10.820 \pm .406$	63.94	$94.81 \pm .003$	$11.296 \pm .396$

butional distance metric. Since we desire a small geometric distance between two samples to produce a large importance weight, the unnormalized weight w_i can be defined by an inverse relation of the form $w_i(\boldsymbol{x}_{n+1}) = 1/(1+D^p)$, where p denotes an additional parameter modulating the slope or skewdness of the weighting distribution. A final weight $\tilde{w}_i$ is then acquired by subsequent normalization.

4 EXPERIMENTS

We next empirically validate our three different approaches to integrating geometric information with conformal procedures. We briefly outline our experiment design, with further details and results in the Appendix. Our code is publicly available at https://github.com/computri/geometric_cp.

Experimental design. As outlined in § 2.3, the canonicalizer is usually trained using a joint task and prior regularization loss $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \beta \cdot \mathcal{L}_{\text{prior}}$. In accordance with conformal prediction we desire a fully *post-hoc* approach amenable to pretrained predictors, and as such leverage canonicalizers trained exclusively with the canonicalization prior via Eq. 7. Consequently, the predictor in our experiments is *pretrained and frozen* when used in conjunction with the CN, whereas relevant data augmentation and equivariant baselines require $\hat{f}_\theta$ to be trained from scratch. We emphasize that our approach leverages the exact same prediction model $\hat{f}_\theta$, pretrained without augmentations.

Given various classification tasks, we employ the popular *Adaptive Predictive Sets* (APS) [Romano et al., 2020b] as our default nonconformity scoring approach for any conformal procedure, and report results for an alternative scoring method by Sadinle et al. [2019] in Appendix B. Following standard practice we report *empirical coverage* and *mean set size* as our metrics to assess the quality of uncertainty estimates [Shafer and Vovk, 2008, Angelopoulos et al., 2024b]. Empirical coverage determines the *validity* of our guarantees by comparing to the target coverage level $(1-\alpha)$, whereas prediction set sizes assess the *efficiency* of the method, and lower set sizes are more informative. We dub the proposed approach CP^2, for the combined use of the canonicalization prior (CP) with conformal prediction.

4.1 ROBUSTNESS TO GEOMETRIC DATA SHIFTS

We assess the method's robustness to three geometric shifts caused by $C4$, $C8$, and $SO(3)$ rotation groups, and across three datasets (CIFAR-10, CIFAR-100, and ModelNet-40) and two data modalities (images and point clouds).

Image classification. We evaluate two ResNet-50 predictors on CIFAR-10 and CIFAR-100 samples subjected to $C4$ and $C8$ rotation shifts. These groups form discretized subgroups of $SO(2)$ with four and eight equidistant elements, respectively. Three model training configurations are considered: *(i)* the prediction models trained in a default, non-augmented manner; *(ii)* the same predictors trained with relevant data augmentations to obtain approximate invariance; and *(iii)* pretrained and frozen predictors with 'bolt-on' canonicalization models trained for $G = 4$ and $G = 8$ group elements. Each configuration is subsequently combined with standard SCP to provide prediction sets with a target coverage rate of $(1-\alpha) = 95\%$.

Classification accuracy and conformal results for CIFAR-100 are given in Tab. 3 (see Tab. 7 for CIFAR-10). For non-shifted data, performance remains comparable. While the base predictor exhibits highest accuracy in that setting, it lacks generalizability under geometric shift, reflected by its poor performance and uninformative set sizes. In contrast, both data-augmented and canonicalization approaches ensure robustness to the shift, while achieving similar accuracy as in the non-shifted setting unless the learned group is misspecified (*i.e.*, trained for $C4$ but exposed to $C8$). We observe that in the inverse case robustness continues to hold, thus suggesting to favour a broader group definition when faced with the risk of unknown group elements. That is, ideally the learned group is chosen to be maximal within constraints on computational resources and accuracy requirements, since a coarser discretization will induce more

Table 4: Results with APS on ModelNet-40 for target coverage $(1-\alpha) = 95\%$ across $SO(3)$ rotation shift. Target coverage is efficiently maintained when no shift occurs, but grows excessively large and uninformative ( ) when the underlying predictor $\hat{f}_{\boldsymbol{\theta}}$ is not equivariant. Results are reported across $T = 10$ random calibration/test splits. # Param. indicates the required number of training parameters. (*) denotes the use of a slightly different data preprocessing and training split.

		No Shift			$SO(3)$ **Rotation Shift**		
Model	# Param.	Acc	Coverage	Set Size	Acc	Coverage	Set Size
PointNet	0.7 M	87.49	$94.93 \pm .008$	$2.133 \pm .079$	8.73	$98.01 \pm .003$	$38.835 \pm .062$
DGCNN	1.8 M	91.41	$94.76 \pm .007$	$1.36 \pm .055$	15.15	$95.62 \pm .004$	$36.218 \pm .161$
Rapidash*	1.7 M	86.51	$95.34 \pm .007$	$1.540 \pm .073$	12.84	$99.95 \pm .000$	$39.956 \pm .018$
PointNet + $SO(3)$	0.7 M	57.46	$95.3 \pm .008$	$6.532 \pm .384$	55.63	$94.85 \pm .011$	$7.058 \pm .497$
DGCNN + $SO(3)$	1.8 M	86.55	$95.18 \pm .004$	$1.508 \pm .040$	85.74	$94.71 \pm .010$	$1.554 \pm .105$
Rapidash* + $SO(3)$	1.7 M	76.18	$94.99 \pm .006$	$3.630 \pm .181$	76.70	$94.72 \pm .005$	$3.566 \pm .151$
Invariant Rapidash* + $SO(3)$	1.7 M	74.39	$95.24 \pm .003$	$4.747 \pm .142$	74.02	$95.26 \pm .007$	$5.255 \pm .348$
Equivariant Rapidash* + $SO(3)$	2.0 M	88.70	$94.82 \pm .006$	$1.427 \pm .040$	87.68	$94.94 \pm .007$	$1.438 \pm .067$
CP^2 (PointNet)	1.3 K	62.11	$95.47 \pm .008$	$6.912 \pm .230$	62.07	$94.94 \pm .007$	$6.604 \pm .226$
CP^2 (DGCNN)	1.3 K	85.78	$95.11 \pm .007$	$2.317 \pm .112$	85.86	$94.46 \pm .008$	$2.262 \pm .105$

discretization artifacts. Overall, our results highlight canonicalization as a light-weight alternative to ensure robustness without necessitating retraining.

Point cloud classification. Unlike 2D images, point clouds exist within a continuous 3D space where rotational shifts are more intrinsic. We evaluate the performance of popular point cloud classifiers PointNet [Qi et al., 2017] and DGCNN [Wang et al., 2019] with and without canonicalization, along with Rapidash [Vadgama et al., 2025], a recent proposal which permits adjustable levels of equivariance—from non-equivariant to fully equivariant. Our results in Tab. 4 echo those from the image domain, revealing that unadjusted base models fail to maintain robustness against orientation shifts in point clouds, resulting in inflated conformal metrics. Conversely, models equipped with data augmentation or equivariance properties demonstrate better resilience to these geometric shifts. In particular this includes canonicalization, which in this particular instance trains a network by *multiple magnitudes* smaller than other approaches (see § A.1 for architecture details). In addition, data augmentations become substantially more expensive to incorporate due to the high degrees of freedom offered by 3D spatial rotations.

4.2 DIAGNOSTICS FOR CONDITIONAL COVERAGE

Next, we leverage the geometric information obtained from the canonicalization network's sample-wise group distributions to construct partition-conditional group distributions $\hat{P}_{G|k}$ following Eq. 8, and visualize the obtained 'group maps' for CIFAR-10 in Fig. 1. In each column, we display the true group distribution $P_{G|k}$—tractable by manually inducing different partition-conditional shifts—and the recovered distribution $\hat{P}_{G|k}$ using the CN. Indeed, we find

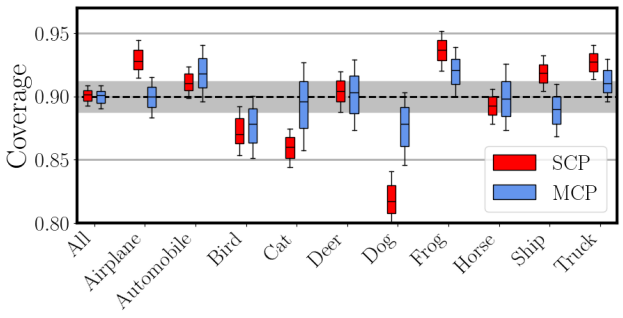
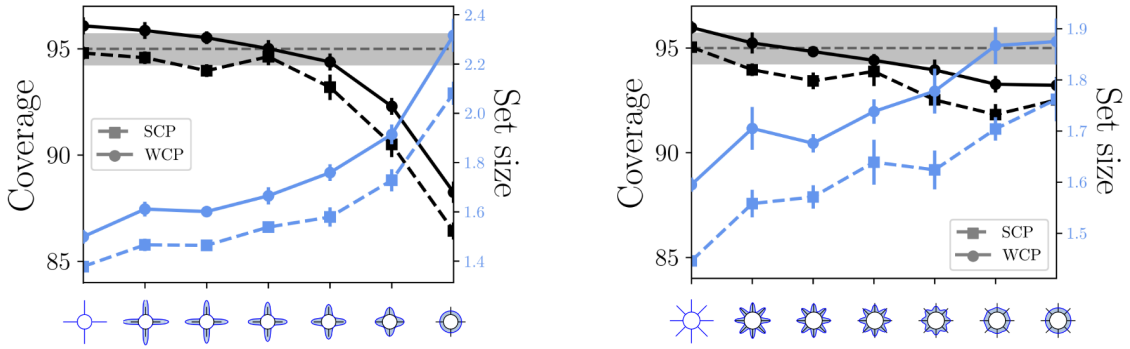


Figure 2: Per-class coverage results for split (SCP) and mondrian conformal prediction (MCP) on CIFAR-10, for a target coverage of $(1-\alpha) = 90\%$. Leveraging MCP on the group partition, improved coverage balance is obtained *by proxy* for the class partition due to an accurately captured geometric relationship between class labels and $C8$ group elements by the group maps (Fig. 1, second column).

that the model can effectively uncover meaningful geometric patterns when particular shifts are imbued on the data. We also visualize the class-conditional group map on the data *without* any geometric shift (Fig. 1, first column) and observe how samples across all classes are predominantly mapped into the identity element, *i.e.* upright. We can interpret the approach as a visual test for exchangeability, assessing whether all bucketed samples across the partition adhere to the same geometric properties (as is the case here).

Additionally, we may determine that particular partitions correlate with particular group elements, and in such cases leverage sample assignments to each entry $\hat{P}_{g|k}$ as an unsupervised proxy for mondrian conformal prediction. While we conformalize directly for the group partition (see Fig. 6), the captured geometric relationship will *by proxy* lead to



(a) $C4$ to $SO(2)$. (b) $C8$ to $SO(2)$.

Figure 3: Coverage and set sizes for the double shift setting (Tab. 2, third row) on CIFAR-10, for split (SCP) and geometrically weighted conformal prediction (WCP). We train the canonicalization network on $\mathcal{D}_{cal}$ with $C4$ resp. $C8$ group knowledge, and gradually induce a secondary test shift by interpolating from benign to strong shift on $SO(2)$ (left to right end of the x-axis showing group distribution). Eventual coverage breakdown is inevitable, but geometric weighting can delay the effect.

more balanced coverage for the associated data partition. We demonstrate this for the class-conditional case in Fig. 2, where MCP applied to the $C8$ group elements—which exhibit a correspondence with CIFAR-10 class labels—substantially improves per-class coverage over SCP due to a better-tailored conformal quantile estimate. Naturally, the proxy relationship is limited by the extent to which a partition-conditional group pattern intersects with multiple partitions, and at the other end no improvements are obtained when partitions exhibit identical group distributions (Fig. 7, left).

4.3 WEIGHTING FOR DOUBLE SHIFT SETTINGS

Finally, we evaluate the double shift setting described in § 3.1 and Tab. 2 (third row). Fig. 3 depicts the encounter of $C4$ and $C8$ discrete rotation shifts on $\mathcal{D}_{cal}$, with a gradual interpolation for different continuous $SO(2)$ shifts (by adjusting data sampling probabilities) on $\mathcal{D}_{test}$. Thus, the secondary test shift ranges from benign (*i.e.* no new group elements) to severe, where the uniform $SO(2)$ group places much probability mass on previously unknown rotations. We employ our geometric weighting scheme in conjunction with weighted conformal prediction, and compare against the standard variant. We clearly observe a coverage breakdown with growing geometric difference between data partitions, since the necessary exchangeability condition between $\mathcal{D}_{cal}$ and $\mathcal{D}_{test}$ is invalidated. Yet, geometric weighting can help delay the effect at the cost of enlarged set sizes, suggesting partial group knowledge can be beneficial to robustness even under *unkown* group actions. However, improvements remain bottlenecked by the static training performance of the canonicalizer, and a more practical deployment should consider an updating step to incorporate new geometric information upon arrival.

5 RELATED WORK

Model-agnostic equivariance and canonicalization. Group equivariance in deep learning models is typically realized through architectures that inherently incorporate equivariance constraints, and as such a wide array of equivariant layers exist [Cohen and Welling, 2016, Weiler and Cesa, 2019, Bekkers, 2020, Cohen et al., 2019, Finzi et al., 2020, 2021, Ruhe et al., 2023]. These models generally require a full training procedure and carefully tailored architecture design. In contrast, a novel range of approaches enable model-agnostic equivariance, integrating it into pretrained backbones with minimal training or finetuning. These methods generally fall into three categories. First, symmetrization methods apply group averaging operators to non-equivariant base models to enforce equivariance, as explored by [Basu et al., 2023b,a, Kim et al., 2023]. Second, frame-averaging techniques, such as those discussed in [Puny et al., 2021, Duval et al., 2023, Atzmon et al., 2021], focus on identifying efficient yet expressive subsets of groups for averaging. The third category encompasses canonicalization methods [Kaba et al., 2023, Mondal et al., 2023, Panigrahi and Mondal, 2024], which offer a competitive and resource-efficient alternative to the first two. Unlike symmetrization and frame-averaging, canonicalization utilizes an auxiliary network to provide explicit per-sample group estimates.

Conformal prediction under distribution shift. A substantial body of recent work has explored the handling of non-exchangeable data sequences, including time series by tracking and adapting miscoverage rates [Gibbs and Candès, 2021, Angelopoulos et al., 2024d, Zaffran et al., 2022, Angelopoulos et al., 2024a] or employing different weighting strategies [Barber et al., 2023, Guan, 2023, Amoukou and Brunel, 2023]. Efforts to directly address shift settings have considered covariate shift [Tibshirani et al., 2019], label shift [Podkopaev and Ramdas, 2021], and broader gener-

alizations [Prinster et al., 2024], with various likelihood ratio-based weights. However, to the best of our knowledge no specific handling of *geometric* shifts has been explored in the literature. The only work explicitly utilizing geometry for conformal prediction is Kaur et al. [2022], who employ a notion of equivariance to detect out-of-distribution samples.

6 DISCUSSION

We propose leveraging geometric information to supplement conformal prediction, robustifying the procedure against *geometric* data shifts and ensuring fundamental conditions such as exchangeability are preserved. We explore multiple applications on integrating the approaches: mitigating performance drops due to geometric variations (§ 4.1), employing it as a diagnostics tool for conditional coverage (§ 4.2), and as a weighting mechanism in settings involving multiple shifts (§ 4.3). While we instantiate our approach using the canonicalization principle, the underlying methodology is broadly applicable and should extend to any sample-wise geometry extractor providing similar group information.

Limitations and Outlook. Our work predominantly explores shifts caused by rotation groups, following Mondal et al. [2023], Kaba et al. [2023] and related works. While theoretically extendable to other groups like roto-reflections, practical implementation across broader groups remains unexplored. Indeed, other research has highlighted the challenges with continuous canonicalization [Dym et al., 2024], aligning with practical difficulties observed by [Mondal et al., 2023] in its application within the image domain. Alternatively, CP could be combined with symmetrization or frame averaging techniques, of which variants exists that use group weighting mechanisms [Dym et al., 2024, Kim et al., 2023]. Such weighting could potentially be used as an alternative to extract per-sample group estimates.

In § 4.3, generalization to unseen group elements assumes probability mass is concentrated on nearby elements, providing relevant information. While often valid, per-sample inaccuracies in the canonicalization network's group predictions (see *e.g.* Tab. 5) can still negatively impact conformal results. More generally, the inclusion of a *trained* geometry extractor as an additional 'bolt-on' module means the pipeline is susceptible to errors propagating down-stream, but can also greatly benefit from future improvements in that regard. Future work can also explore the use of geometric information for more conformal prediction settings including regression tasks [Sesia and Romano, 2021b], stream data [Gibbs and Candès, 2021] or more general risk notions [Angelopoulos et al., 2024c]. A more in-depth exploration of WCP could help develop robuster geometric weighting schemes, perhaps through Prinster et al. [2024]'s lens as weighted permutations. Ultimately, the intersection of conformal prediction and geometric deep learning remains largely unexplored, offering promising directions for future work.

Author Contributions

PvdL and AT contributed jointly to ideation and methodology, and co-authored the paper. In addition, PvdL initiated the preliminary approach and conducted all experiments, while AT initiated the theoretical motivation. EB contributed to ideation development, project guidance and feedback.

Acknowledgements

We thank Siba Smarak Panigrahi for valuable discussions about the canonicalization codebase upon which this work relies. We also thank Rajeev Verma and Sharvaree Vadgama for insightful discussions on geometric uncertainty.

References

James Urquhart Allingham, Bruno Kacper Mlodozeniec, Shreyas Padhy, Javier Antorán, David Krueger, Richard E. Turner, Eric Nalisnick, and José Miguel Hernández-Lobato. A generative model of symmetry transformations. *Advances in Neural Information Processing Systems*, 2024.

Salim I Amoukou and Nicolas JB Brunel. Adaptive conformal prediction by reweighting nonconformity score. *arXiv Preprint (arXiv:2303.12695)*, 2023.

Anastasios Angelopoulos, Emmanuel Candes, and Ryan J Tibshirani. Conformal pid control for time series prediction. *Advances in Neural Information Processing Systems*, 2024a.

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 2023.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv Preprint (arXiv:2411.11824)*, 2024b.

Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *International Conference on Learning Representations*, 2024c.

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. Online conformal prediction with decaying step sizes. *International Conference on Machine Learning*, 2024d.

Matan Atzmon, Koki Nagano, Sanja Fidler, Sameh Khamis, and Yaron Lipman. Frame averaging for equivariant shape space learning. *Conference on Computer Vision and Pattern Recognition*, 2021.

Rina Foygel Barber, Emmanuel J Candes, Aaditya Ramdas, and Ryan J Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 2023.

Sourya Basu, Pulkit Katdare, Prasanna Sattigeri, Vijil Chenthamarakshan, Katherine Driggs-Campbell, Payel Das, and Lav R. Varshney. Efficient equivariant transfer learning from pretrained models. *Advances in Neural Information Processing Systems*, 2023a.

Sourya Basu, Prasanna Sattigeri, Karthikeyan Natesan Ramamurthy, Vijil Chenthamarakshan, Kush R. Varshney, Lav R. Varshney, and Payel Das. Equi-tuning: Group equivariant fine-tuning of pretrained models. *The Thirty-Seventh AAAI Conference on Artificial Intelligence*, 2023b.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 2019.

Erik J Bekkers. B-spline cnns on lie groups. *International Conference on Learning Representations*, 2020.

Benjamin Bloem-Reddy, Yee Whye, et al. Probabilistic symmetries and invariant neural networks. *Journal of Machine Learning Research*, 2020.

Johann Brehmer, Sönke Behrends, Pim de Haan, and Taco Cohen. Does equivariance matter at scale? *arXiv Preprint (arXiv:2410.23179)*, 2024.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv Preprint (arXiv:2104.13478)*, 2021.

Maxime Cauchois, Suyash Gupta, and John C Duchi. Knowing what you know: Valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, 2021.

Taco Cohen and Max Welling. Group equivariant convolutional networks. *International Conference on Machine Learning*, 2016.

Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral CNN. *International Conference on Machine Learning*, 2019.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J. Guibas. Vector neurons: A general framework for so(3)-equivariant networks. *International Conference on Computer Vision*, 2021.

Alexandre Agm Duval, Victor Schmidt, Alex Hernández-García, Santiago Miret, Fragkiskos D. Malliaros, Yoshua Bengio, and David Rolnick. FAENet: Frame averaging equivariant GNN for materials modeling. *International Conference on Machine Learning*, 2023.

Nadav Dym, Hannah Lawrence, and Jonathan W. Siegel. Equivariant frames and the impossibility of continuous canonicalization. *International Conference on Machine Learning*, 2024.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Marc Finzi, Max Welling, and Andrew Gordon Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. *International Conference on Machine Learning*, 2021.

Huazhu Fu, Yitian Zhao, Pew-Thian Yap, Carola-Bibiane Schönlieb, and Alejandro F Frangi. Guest editorial special issue on geometric deep learning in medical imaging. *IEEE Transactions on Medical Imaging*, 2023.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 2023.

Isaac Gibbs and Emmanuel J. Candès. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 2021.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 2023.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch Multivalid Conformal Prediction. *International Conference on Learning Representations*, 2023.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. *International Conference on Machine Learning*, 2023.

Ramneet Kaur, Susmit Jha, Anirban Roy, Sangdon Park, Edgar Dobriban, Oleg Sokolsky, and Insup Lee. idecode: In-distribution equivariance for conformal out-of-distribution detection. *AAAI Conference on Artificial Intelligence*, 2022.

Abbas Khosravi, Saeid Nahavandi, Doug Creighton, and Amir F Atiya. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on Neural Networks*, 2011.

Jinwoo Kim, Tien Dat Nguyen, Ayhan Suleymanzade, Hyeokjun An, and Seunghoon Hong. Learning probabilistic symmetrization for architecture agnostic equivariance. *Advances in Neural Information Processing Systems*, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. *International Conference on Machine Learning*, 2021.

Benjamin Kompa, Jasper Snoek, and Andrew L Beam. Empirical frequentist coverage of deep learning uncertainty quantification procedures. *Entropy*, 2021.

Spyros Makridakis and Nikolas Bakas. Forecasting and uncertainty: A survey. *Risk and Decision Analysis*, 2016.

Arnab Kumar Mondal, Siba Smarak Panigrahi, Sékou-Oumar Kaba, Sai Rajeswar, and Siamak Ravanbakhsh. Equivariant adaptation of large pretrained models. *Advances in Neural Information Processing Systems*, 2023.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 2019.

Lace MK Padilla, Maia Powell, Matthew Kay, and Jessica Hullman. Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations. *Frontiers in Psychology*, 2021.

Siba Smarak Panigrahi and Arnab Kumar Mondal. Improved canonicalization for model agnostic equivariance. *Conference on Computer Vision and Pattern Recognition (EquiVision Workshop)*, 2024.

Harris Papadopoulos, Volodya Vovk, and Alex Gammerman. Conformal prediction with Neural Networks. *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2007.

Aleksandr Podkopaev and Aaditya Ramdas. Distribution-free uncertainty quantification for classification under label shift. *Uncertainty in Artificial Intelligence*, 2021.

Drew Prinster, Samuel Don Stanton, Anqi Liu, and Suchi Saria. Conformal validity guarantees exist for any data distribution (and how to find them). *International Conference on Machine Learning*, 2024.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Edward J. Smith, Ishan Misra, Aditya Grover, and Yaron Lipman. Frame averaging for invariant and equivariant network design. *International Conference on Learning Representations*, 2021.

Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *Conference on Computer Vision and Pattern Recognition*, 2017.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. MIT Press, 2022.

Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel J Candès. With Malice Towards None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review*, 2020a.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 2020b.

David W. Romero and Suhas Lohit. Learning partial equivariances from data. *Advances in Neural Information Processing Systems*, 2022.

David Ruhe, Johannes Brandstetter, and Patrick Forré. Clifford group equivariant neural networks. *Advances in Neural Information Processing Systems*, 2023.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 2019.

Matteo Sesia and Yaniv Romano. Conformal Prediction using Conditional Histograms. *Advances in Neural Information Processing Systems*, 2021a.

Matteo Sesia and Yaniv Romano. Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems*, 2021b.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 2008.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 2019.

Paolo Toccaceli and Alexander Gammerman. Combination of inductive mondrian conformal predictors. *Machine Learning*, 2019.

Alonso Urbano and David W. Romero. Self-supervised detection of perfect and partial input-dependent symmetries. *International Conference on Machine Learning (GRaM Workshop)*, 2024.

Sharvaree Vadgama, Mohammad Mohaiminul Islam, Domas Buracus, Christian Shewmake, and Erik Bekkers. On the utility of equivariance and symmetry breaking in deep learning architectures on point clouds. *arXiv Preprint (arXiv:2501.01999)*, 2025.

Putri A. van der Linden, Alejandro García-Castellanos, Sharvaree Vadgama, Thijs P. Kuipers, and Erik J. Bekkers. Learning symmetries via weight-sharing with doubly stochastic tensors. *Advances in Neural Information Processing Systems*, 2024.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *Association for Computing Machinery*, 2019.

Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. *Advances in Neural Information Processing Systems*, 2019.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *International Conference on Learning Representations*, 2023.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. *International Conference on Machine Learning*, 2022.

CP^2: Leveraging Geometry for Conformal Prediction via Canonicalization — Supplementary Material —

CONTENTS

1 Introduction **1**

2 Background **2**

2.1 Conformal Prediction . . . 2

2.2 Group Equivariance and Invariance . . . 2

2.3 Equivariance via Canonicalization . . . 3

3 Geometric Information for Conformal Prediction **3**

3.1 Use Cases for Conformal Prediction . . . 4

4 Experiments **6**

4.1 Robustness to Geometric Data Shifts . . . 6

4.2 Diagnostics for Conditional Coverage . . . 7

4.3 Weighting for Double Shift Settings . . . 8

5 Related Work **8**

6 Discussion **9**

A Additional Experiment Details **14**

A.1 Robustness to Geometric Data Shifts . . . 14

A.2 Diagnostics for Conditional Coverage . . . 14

A.3 Weighting for Double Shift Settings. . . . 15

B Additional Experiment Results **16**

A ADDITIONAL EXPERIMENT DETAILS

Considerations on the canonicalization network. It is important to stress the motivation of a light-weight, *post-hoc* method when making architectural design choices about the canonicalization network (CN). As such we desire an efficient and thus smaller and potentially less expressive CN. Naturally, a small performance drop over the symmetry-unaware predictor is expected by incorporating the CN, since as a trained model it is prone to (some) pose prediction errors itself. Additionally, both the limited expressivity of the light-weight CN as well as the resolution on $SO(2)$ induces discretization artifacts. We examine the CN's miscanonicalization in Tab. 5, reporting the fraction of correctly predicted group elements on CIFAR-10 data subject to $C4$ and $C8$ rotations. We additionally nuance that the miscanonicalization results do not directly translate to down-stream lower prediction accuracy, as the predictor itself can exhibit robustness to minor pose variations (*e.g.* correctly classifying images with rotation angles in $[-45°, 45°]$) and hence may still correctly predict a target despite erroneous pose alignment.

Table 5: Fraction of group elements correctly predicted by trained canonicalization networks on CIFAR-10. We subsequently evaluate the models on a hold-out split of data under the same group effects, *i.e.* $C4$ or $C8$.

Model	**% Corr. angles**
Canonicalization with $(G = 4)$	87.46
Canonicalization with $(G = 8)$	87.23

A.1 ROBUSTNESS TO GEOMETRIC DATA SHIFTS

Image canonicalization network. For the image domain, when taking into account the aforementioned desire for an efficient geometric module, we restrict ourselves to $C4$ and $C8$-equivariant canonicalization networks. We adopt the models described in Mondal et al. [2023], employing a compact 3-layer, G-equivariant WideResNet, where G-equivariance is achieved through the use of E2CNN [Weiler and Cesa, 2019]. All models are trained for a maximum of 100 epochs with early stopping, and optimized using Adam.

Point cloud canonicalization network. For the continuous point cloud domain, we similarly follow the approaches outlined in Mondal et al. [2023] by adopting a compact Vector Neuron model [Deng et al., 2021]. These models are trained for 250 epochs with a cosine learning rate scheduler, and optimized using Adam.

A.2 DIAGNOSTICS FOR CONDITIONAL COVERAGE

Intuition. We explore the hypothesis that different group elements (*e.g.* particular rotation angles) can correlate with specific data partitions due to their distinct geometric properties [Urbano and Romero, 2024, van der Linden et al., 2024, Allingham et al., 2024, Romero and Lohit, 2022]. For example, isotropic shapes such as a ring or the digit "0" (*e.g.* in MNIST) may withstand arbitrary rotations without altering their class identity or losing significant visual features. Hence their geometric pose may be naturally uniformly distributed over the rotation group. Conversely, shapes such as the digit "6" transform into a "9" when rotated at 180°, potentially leading to erroneous prediction. Consequently, one would not expect to observe such group elements to meaningfully contribute to the shape's natural pose distribution. Our experiments in § 4.2 manually induce such shifts (*e.g.* on class labels) in order to highlight the canonicalization network's accurate recovery of such geometric behaviour.

Experimental design. We induce several group shifts conditioned on particular target partitions:

- `dirac`: A dirac distribution over the group, pinpointing a single group element per partition;
- `normal`: A normal distribution over the group; and
- `var-gauss`: various Gaussian distributions with standard deviations in $[0.0001, 0.001, 0.01, 0.1, 1.0, 10.0]$.

To improve the visual recovery of partition-conditional group effects, we additionally exclude data points for which the canonicalization network's predicted group probability falls below a predefined threshold, ensuring that only samples with confident group predictions are taken into account. This aids in counteracting some of the canonicalization's erroneous predictions (see Tab. 5) to better demonstrate why mondrian conformal prediction may be useful when clear patterns exist.

A.3 WEIGHTING FOR DOUBLE SHIFT SETTINGS.

Intuition. The canonicalization network's role is to mitigate the first shift between $\mathcal{D}_{train}$ and $\mathcal{D}_{cal}$, but in the double-shift setting of § 4.3 we also encounter a subsequent shift between $\mathcal{D}_{cal}$ and $\mathcal{D}_{test}$ (see Tab. 2, third row), such as from known discrete group elements in $C8$ to potentially any rotation in $SO(2)$ (see Fig. 3, right). From empirical observations and prior studies, minor rotations (e.g., within ± 5 degrees) have shown to enhance the accuracy of down-stream pose prediction tasks. This improvement is often attributed to the alignment with natural object variations captured in datasets [Mondal et al., 2023]. This insight suggests that within small deviations from known group elements, a well-trained CN is capable of accurately identifying the nearest group element. This accuracy decreases as the continuous rotation deviates further from these discretized elements, reaching maximum ambiguity at positions equidistant from two neighboring group elements (*i.e.* maximal shift). Therefore, when a test sample's transformation is close to one of these discretized rotations, the CN tends to assign higher probabilities to that group element or its immediate neighbors. We harness these insights on the CN's probabilistic output to obtain geometry-informed weights for weighted conformal prediction, enhancing robustness to rotations not explicitly covered by the discrete, known group elements.

Experimental design. To navigate the transition between the (exchangeable) discrete setting and the (non-exchangeable) uniform $SO(2)$ group, we model a group distribution on the sphere. Specifically, we define it as either discrete peaks at the $C4$ or $C8$ elements, or as a continuous distribution using a mixture of 'von Mises' distributions each centered at $C4$ or $C8$ elements. The 'von Mises' p.d.f. is of the form $f(x \mid \mu, \kappa) = 1/(2\,\pi\, I_0(\kappa)) \cdot \exp(\kappa \cos(x))$, where μ denotes a location parameter and κ controls the concentration of mass around μ. Varying κ facilitates the interpolation between discrete $C4$ or $C8$ sampling and more uniform $SO(2)$ sampling. In Fig. 3, we use $\kappa = [50, 40, 30, 20, 10]$ as interpolative factors, and visualize the resulting spherical group distributions on the x-axis. Regarding the inverse geometric weighting relationship described in § 3.1, we visualize different values of the modulating parameter p and their effect on the weighting distribution in Fig. 4; and their impact on the $C4$ to $SO(2)$ double-shift setting in Fig. 5. For the main paper, we opt for a cross-entropy distance metric and set $p = 2.0$ as it empirically displays a good trade-off between set size and coverage target.

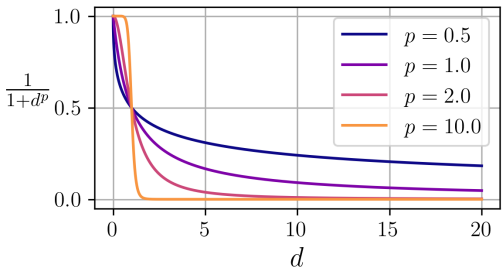


Figure 4: Illustration of the impact of the modulation parameter p on the inverse geometric weighting relation. As p increases, the weighting distribution approaches a binary mask for smal distances.

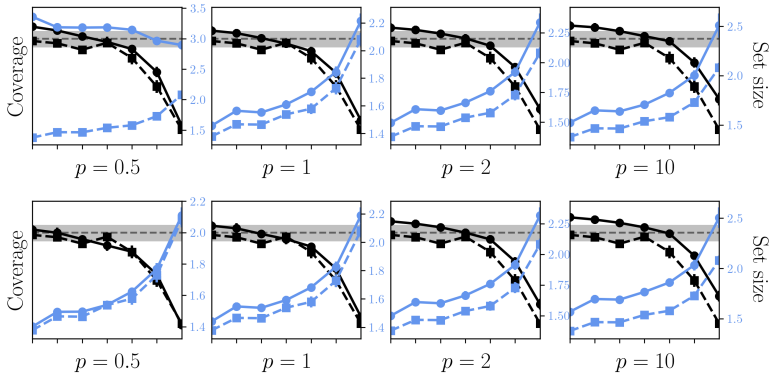


Figure 5: Ablation on the modulation parameter p and two different distributional distance metrics (*top*: KL-divergence, *bottom*: cross-entropy) for the double-shift setting (§ 4.3) for $C4$ to $SO(2)$.

B ADDITIONAL EXPERIMENT RESULTS

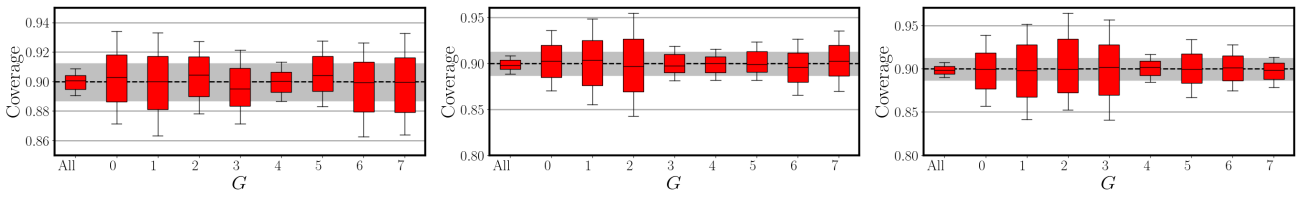


Figure 6: Empirical coverage for mondrian conformal prediction on the exact geometric group partitions for shift settings 2, 3 and 4 in Fig. 1. As we perform MCP directly on these partitions, per-group target coverage is nominally guaranteed.

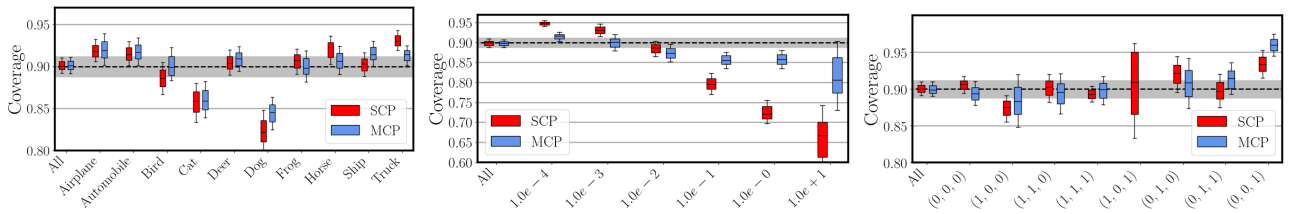


Figure 7: Empirical coverage for the target partitions via split (SCP) and mondrian conformal prediction (MCP) for shift settings 1, 3 and 4 in Fig. 1. Improved coverage balance is obtained *by proxy* when a strong and accurately caputed geometric relationship is apparent. *Left*: SCP and MCP match when all target partitions exhibit identical group distributions, *middle, right*: MCP yields more balanced coverage in the presence of a meaningful *by proxy* relation.

Robustness to geometric shift with Thr [Sadinle et al., 2019]. Accuracy and conformal results for the geometric shift experiments from § 4.1 on both images and point clouds using another conformal scoring approach (Thr [Sadinle et al., 2019]). This conformal scoring function is simply defined as $s(\boldsymbol{x}_i, \boldsymbol{y}_i) = 1 - \hat{p}(\mathbf{y}_i = \boldsymbol{y}_i | \boldsymbol{x}_i)$ for any true class label $\boldsymbol{y}_i$. The interpretation of obtained results is consistent with those in the main paper using *APS* [Romano et al., 2020b].

Table 6: Results with **Thr [Sadinle et al., 2019] on ModelNet-40** for target coverage $(1-\alpha) = 95\%$ across $SO(3)$ rotation shift. Target coverage is efficiently maintained when no shift occurs, but grows excessively large and uninformative ( ) when the underlying predictor $\hat{f}_{\boldsymbol{\theta}}$ is not equivariant. Results are reported across $T = 10$ random calibration and test splits. # Param. indicates the required number of trained parameters. (*) Use slightly different train/test splits and data preprocessing.

		No Shift			$SO(3)$ **Rotation Shift**		
Model	# Param.	Acc	Cov	Set Size	Acc	Cov	Set Size
PointNet	0.7 M	87.49	$95.11 \pm .006$	$1.368 \pm .019$	8.73	$100.00 \pm .000$	$40.000 \pm .000$
DGCNN	1.8 M	91.41	$94.68 \pm .005$	$1.103 \pm .015$	15.15	$100.00 \pm .000$	$40.000 \pm .000$
Rapidash*	1.7 M	86.59	$95.12 \pm .010$	$1.498 \pm .078$	12.12	$100.00 \pm .000$	$40.000 \pm .000$
PointNet + $SO(3)$	0.7 M	57.46	$95.3 \pm .008$	$6.532 \pm .384$	55.63	$94.85 \pm .011$	$7.058 \pm .497$
DGCNN + $SO(3)$	1.8 M	86.55	$95.18 \pm .004$	$1.508 \pm .040$	85.74	$94.71 \pm .010$	$1.554 \pm .105$
Rapidash* + $SO(3)$	1.7 M	76.50	$95.07 \pm .005$	$3.519 \pm .212$	75.93	$95.01 \pm .007$	$3.323 \pm .208$
Invariant Rapidash* + $SO(3)$	1.7 M	74.47	$95.00 \pm .008$	$4.525 \pm .368$	74.35	$95.21 \pm .009$	$4.650 \pm .306$
Equivariant Rapidash* + $SO(3)$	2.0 M	88.21	$95.05 \pm .006$	$1.446 \pm .038$	87.60	$95.14 \pm .006$	$1.416 \pm .032$
PRLC-PointNet	1.3 K	62.11	$95.60 \pm .010$	$5.936 \pm .361$	62.07	$95.09 \pm .008$	$5.745 \pm .266$
PRLC-DGCNN	1.3 K	85.78	$95.42 \pm .005$	$1.663 \pm .052$	85.86	$94.55 \pm .012$	$1.594 \pm .083$

Table 7: Results with **APS [Romano et al., 2020b] on CIFAR-10** for target coverage $(1-\alpha) = 95\%$ across different rotation shifts. Target coverage is efficiently maintained when no shift occurs, but grows excessively large and uninformative ( ) when the underlying predictor $f_{\boldsymbol{\theta}}$ is not equivariant, or if the wrong group is specified (controlling for $C4$ but exposed to $C8$). Results are reported across $T = 10$ random calibration and test splits. # Param. indicates the required number of *trained* parameters.

		No Shift			$C4$ **Rotation Shift**			$C8$ **Rotation Shift**		
Model	# Param.	Acc	Cov	Set Size	Acc	Cov	Set Size	Acc	Cov	Set Size
$\hat{f}_\theta$	23.5 M	92.82	$94.94 \pm .004$	$1.137 \pm .005$	52.50	$95.06 \pm .005$	$5.324 \pm .078$	45.39	$95.28 \pm .005$	$6.420 \pm .154$
$\hat{f}_\theta$ + $SO(2)$ Aug.	23.5 M	84.97	$94.82 \pm .004$	$1.560 \pm .019$	85.03	$95.01 \pm .004$	$1.534 \pm .024$	83.89	$94.99 \pm .006$	$1.541 \pm .027$
$\hat{f}_\theta$ + C_4 Aug.	23.5 M	88.08	$95.07 \pm .005$	$1.362 \pm .019$	87.81	$95.28 \pm .003$	$1.356 \pm .022$	70.56	$95.11 \pm .004$	$4.213 \pm .059$
$\hat{f}_\theta$ + C_8 Aug.	23.5 M	86.39	$95.20 \pm .003$	$1.477 \pm .018$	86.41	$95.11 \pm .005$	$1.447 \pm .022$	85.36	$95.11 \pm .005$	$1.459 \pm .025$
CP^2 with G=4	0.25 M	88.09	$94.74 \pm .005$	$1.374 \pm .017$	88.09	$95.10 \pm .004$	$1.387 \pm .017$	66.80	$94.76 \pm .005$	$3.814 \pm .087$
CP^2 with G=8	0.51 M	88.13	$94.93 \pm .003$	$1.369 \pm .015$	88.13	$95.02 \pm .003$	$1.374 \pm .015$	86.97	$94.95 \pm .003$	$1.462 \pm .020$

Table 8: Results with **Thr [Sadinle et al., 2019] on CIFAR-10** for target coverage $(1-\alpha) = 95\%$ across different rotation shifts. Target coverage is efficiently maintained when no shift occurs, but grows excessively large and uninformative ( ) when the underlying predictor $\hat{f}_{\boldsymbol{\theta}}$ is not equivariant, or if the wrong group is specified (controlling for $C4$ but exposed to $C8$). Results are reported across $T = 10$ random calibration and test splits. # Param. indicates the required number of trained parameters.

		No Shift			$C4$ **Rotation Shift**			$C8$ **Rotation Shift**		
Model	# Param.	Acc	Cov	Set Size	Acc	Cov	Set Size	Acc	Cov	Set Size
$\hat{f}_\theta$	23.5 M	92.82	$94.79 \pm .004$	$1.061 \pm .008$	52.50	$95.28 \pm .003$	$5.171 \pm .050$	45.39	$94.50 \pm .006$	$6.187 \pm .114$
$\hat{f}_\theta$ + $SO(2)$ Aug.	23.5 M	85.00	$94.73 \pm .004$	$1.485 \pm .022$	85.03	$95.05 \pm .004$	$1.490 \pm .021$	83.90	$95.01 \pm .004$	$1.476 \pm .020$
$\hat{f}_\theta$ + C_4 Aug.	23.5 M	88.08	$94.92 \pm .005$	$1.304 \pm .024$	87.81	$95.27 \pm .004$	$1.308 \pm .018$	70.56	$95.11 \pm .004$	$4.077 \pm .055$
$\hat{f}_\theta$ + C_8 Aug.	23.5 M	86.39	$95.14 \pm .002$	$1.430 \pm .012$	86.41	$94.97 \pm .005$	$1.402 \pm .025$	85.36	$95.08 \pm .004$	$1.417 \pm .020$
CP^2 with G=4	0.25 M	88.12	$94.75 \pm .003$	$1.340 \pm .012$	88.12	$95.03 \pm .003$	$1.351 \pm .018$	66.84	$94.90 \pm .005$	$3.742 \pm .067$
CP^2 with G=8	0.51 M	88.27	$94.90 \pm .003$	$1.329 \pm .017$	88.27	$94.92 \pm .002$	$1.331 \pm .015$	87.04	$94.78 \pm .003$	$1.417 \pm .017$

Table 9: Results with **Thr [Sadinle et al., 2019] on CIFAR-100** for target coverage $(1-\alpha) = 95\%$ across different rotation shifts. Target coverage is efficiently maintained when no shift occurs, but grows excessively large and uninformative ( ) when the underlying predictor $\hat{f}_{\boldsymbol{\theta}}$ is not equivariant, or if the wrong group is specified (controlling for $C4$ but exposed to $C8$). Results are reported across $T = 10$ random calibration and test splits. # Param. indicates the required number of trained parameters.

		No Shift			$C4$ **Rotation Shift**			$C8$ **Rotation Shift**		
Model	# Param.	Acc	Cov	Set Size	Acc	Cov	Set Size	Acc	Cov	Set Size
$\hat{f}_\theta$	23.7 M	71.66	$95.15 \pm .003$	$5.57 \pm .102$	40.17	$100.00 \pm .000$	$100.000 \pm .000$	33.68	$100.00 \pm .000$	$100.000 \pm .000$
$\hat{f}_\theta$ + $SO(2)$ Aug.	23.7 M	60.13	$95.00 \pm .005$	$10.614 \pm .407$	59.72	$95.06 \pm .005$	$10.895 \pm .383$	58.27	$95.05 \pm .005$	$10.534 \pm .433$
$\hat{f}_\theta$ + C_4 Aug.	23.7 M	63.03	$95.08 \pm .006$	$9.361 \pm .428$	62.72	$95.16 \pm .004$	$9.399 \pm .328$	49.72	$100.00 \pm .000$	$100.000 \pm .000$
$\hat{f}_\theta$ + C_8 Aug.	23.7 M	62.53	$95.35 \pm .005$	$10.063 \pm .424$	62.37	$95.27 \pm .003$	$9.990 \pm .180$	60.82	$95.26 \pm .003$	$9.856 \pm .171$
CP^2 with G=4	1.0 M	65.48	$95.20 \pm .005$	$10.173 \pm .491$	65.48	$95.02 \pm .004$	$9.982 \pm .388$	48.33	$98.46 \pm .024$	78.834 ± 32.332
CP^2 with G=8	2.0 M	65.51	$95.39 \pm .005$	$10.674 \pm .465$	65.51	$94.99 \pm .004$	$10.271 \pm .401$	64.16	$94.96 \pm .004$	$11.10 \pm .312$