Extended Abstract for: CHiQPM: Calibrated Hierarchical Interpretable Image Classification

Thomas Norrenbrock[1] Timo Kaiser[1] Sovan Biswas[2] Neslihan Kose[2] Bodo Rosenhahn[1]
[1]Institute for Information Processing (TNT), L3S - Leibniz Universität Hannover, Germany
[2]Intel Labs, Germany
{norrenbr, kaiser, rosenhahn}@tnt.uni-hannover.de
{sovan.biswas, neslihan.kose.cihangir}@intel.com

Abstract

This extended abstract summarizes our work presenting the Calibrated Hierarchical QPM (CHiQPM), originally accepted at NeurIPS 2025. It extends QPM, a model with built-in global interpretability. Alongside global explanations, detailed local explanations are a crucial complement to effectively support human experts during inference. CHiQPM offers uniquely comprehensive global and local interpretability, paving the way for human-AI complementarity. It achieves superior global interpretability by contrastively explaining the majority of classes and offers novel hierarchical explanations that are more similar to how humans reason and can be traversed to offer a built-in interpretable Conformal prediction (CP) method. Our comprehensive evaluation shows that CHiQPM achieves state-of-the-art accuracy as a point predictor, maintaining 99% accuracy of non-interpretable models. Furthermore, its calibrated set prediction is competitively efficient to other CP methods, while providing interpretable predictions of coherent sets along its hierarchical explanation. The code with demo is published: `https://github.com/ThomasNorr/CHiQPM/`

1. Introduction

Using more transparent models, *e.g*. those that are interpretable by-design, is a promising approach to facilitate safe deployment of deep neural networks and is even required by law for some applications [16]. For domains like autonomous driving, with no expert present during inference, models with built-in global interpretability, that can generally explain their behavior, are valuable as their reasoning can be robustly tested and verified before deployment. QPM and Q-SENN [7, 9] follow that goal by enforcing very compact class representations, that are made up of general, diverse and contrastive features, which are properties of human-friendly explanations [5]. Considering the cognitive limitations of average humans [4], QPM represents every class with the binary assignment of very few, usually ≤ 5, broadly shared features. A consequence is the emergence of highly contrastive class representations. Similar to Figure 1, the difference between two classes in the model's representation space can be concretely pointed out, like differentiating the two birds via their eye, just like humans do [18]. While the contrastive representations are very helpful for QPM, they are also fairly rare, *e.g*.on average just 0.13% pairs per class on CUB-2011 [18].

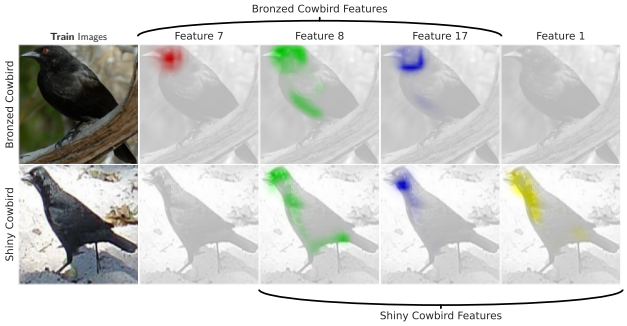

Figure 1. Contrastive global Explanation, comparing the class representations of Shiny and Bronzed Cowbirds for CHiQPM that represents every class with 3 of 30 features. The cowbirds are differentiated based on the red eye.

Other domains, *e.g*.medicine or science, can profit off of additional interpretability. When a human expert is present, they should be supported rather than replaced, a notion known as human-AI complementarity. While global interpretability is still beneficial in this scenario, the value of explaining the decision for a single sample rises, known as local explanation. These typically have the form of saliency maps, such as GradCAM [13], that visualize where the explainee saw support for its decision. They can also be meaningfully computed for individual features if the globally interpretable model can be decomposed into detecting general human un-

derstandable concepts [9]. This crucial property is a key feature of our proposed CHiQPM, as demonstrated in Figures 1 and 2. However, those heatmaps generally do not transport a notion of certainty. Therefore, predicting a set of classes with configurable guarantees on the accuracy using Conformal Prediction (CP) [17] has emerged as a promising direction for supporting experts [14, 15]. Intuitively, more classes are predicted for uncertain or less conform samples, whereas a point-predictor always predicts just one class. However, these sets typically contain a larger variety of classes, that resemble the misalignment between human and machine representations.

This work introduces the Calibrated Hierarchical QPM (CHiQPM). It improves the global interpretability of QPM while maintaining or improving the state-of-the-art accuracy by enforcing more pairs of classes with highly contrastive class representations and adapting the training pipeline to ensure class representations made of interpretable features via the proposed Feature Grounding Loss $\mathcal{L}_{\text{feat}}$ combined with ReLU activation. CHiQPM is the first model with a built-in interpretable set prediction that can be calibrated via CP, inheriting all its robust guarantees. Intuitively, CHiQPM predicts sets of classes by predicting all classes that share the dominant n features with the most likely class, *e.g.* predicting all the black birds in the hierarchical explanation in Figure 2 below the blue feature. This results in a novel way of providing hierarchical local explanations and traversing them to dynamically and understandably construct coherent prediction sets similar to how a human would reason. Considering the graph in Figure 2, the CHiQPM found the green and blue feature, that identify black birds, but no sufficient evidence to differentiate between them. Therefore, it predicts the coherent set of various black birds, including the correct class.

Our main **contributions** are:

- We present the Calibrated Hierarchical QPM (CHiQPM). It is based on a heavily constrained discrete quadratic problem (QP), that selects features from a black-box model and assigns them to classes. The features of CHiQPM then adapt to the optimal solution, resulting in a globally and locally interpretable model.
- CHiQPM offers novel hierarchical local explanations and can be calibrated to reach a target coverage with competitive efficiency while ascending through its dynamically constructed interpretable class hierarchy and selecting the appropriate level. Thus, CHiQPM can be considered an interpretable conformal predictor.
- We present the Feature Grounding Loss $\mathcal{L}_{\text{feat}}$, which, alongside an additional ReLU, leads to learning more grounded and sparser features that facilitate compact hierarchical explanations along more human concepts.
- The state-of-the-art performance of CHiQPM as point- and built-in interpretable calibrated coherent set-predictor is evaluated across datasets, including ImageNet-1K [11], where the gap to the black-box baseline is more than halved.

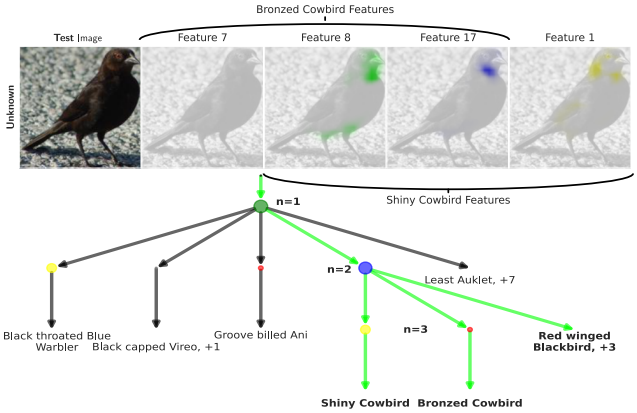

Figure 2. Exemplary local explanation provided by our CHiQPM, with the global explanation in Figure 1, for a difficult test image of a Bronzed Cowbird with a pale red eye that is not clearly visible. This leads to negligible activation of the red-eye detecting Feature 7. The calibrated CHiQPM provides a hierarchical explanation that communicates clear evidence for the predicted coherent set of black birds, but no sufficient evidence to differentiate between them.

2. Method

The proposed CHiQPM classifies an input image by extracting a feature vector $\boldsymbol{f}^* \in \mathbb{R}^{n_f^*}$ and multiplying it by a sparse binary assignment matrix $\boldsymbol{W}^* \in \{0,1\}^{n_c \times n_f^*}$. Compared to QPM [9], we apply an additional ReLU to the features $\mathbf{f}^* = \text{ReLU}(\hat{\mathbf{f}})$, so that CHiQPM only reasons positively, and negligible activations are suppressed. Notably, all classes in CHiQPM are represented with the same number n_{wc} of features per class, which leads to easily comparable class representations. It is easiest, when two classes share exactly $n_{\text{wc}} - 1$ features, as the differentiating factor can then be concretely pointed out, as shown in Figure 1. We denote the set of these pairs with

$$\mathbb{P} = \{(i,j) : 1 \leq i < j \leq n_c \text{ and } \boldsymbol{w}_i^* \boldsymbol{w}^{*T}_j = n_{\text{wc}} - 1\} \quad (1)$$

CHiQPM follows a similar pipeline to QPM [9]. CHiQPM improves upon QPM with easier interpretable class representations for more of the classes via an increased $|\mathbb{P}|$, directly enforced in the QP (Sec. 2.1), alongside a built-in interpretable conformal prediction method traversing the novel local hierarchical explanations (Sec. 2.2) and an improved fine-tuning with a new Feature Grounding Loss $\mathcal{L}_{\text{feat}}$ that improves grounding and compactness of CHiQPM's explanations (Sec. 2.3).

2.1. Hierarchical Constraint

In order to ensure a higher cardinality $|\mathbb{P}|$ resulting in more easily comparable class representations, an additional constraint is added to the quadratic problem. A set $\mathbb{K}$ of pairs of classes that are highly similar in our dense model is determined, and we ensure that they are incorporated into the resulting $\mathbb{P}$. $\mathbb{K}$ contains the $\rho \cdot n_c$ most similar class pairs based on the dense model's class-class similarity matrix used in the QP, where ρ is a density hyperparameter. Representing these $\mathbb{K}$ class pairs similarly is enforced in the QP:

$$(\boldsymbol{w}_c \circ \boldsymbol{w}_{c'})^T \boldsymbol{s} = n_{\text{wc}} - 1 \quad \forall (c, c') \in \mathbb{K} \tag{2}$$

A higher ρ increases the number of classes that can be contrastively globally explained as in Figure 1, thus improving global interpretability. Forcing similar classes to be represented very similarly by CHiQPM further induces an efficient use of the n_f^* features, as shared concepts need only be detected by shared features.

2.2. Hierarchical Explanation

Our CHiQPM enables the construction of hierarchical local explanations for a concrete test sample, as shown in Figure 2. It contains nodes for all nonzero feature activations and indicates the presence of all classes that are assigned to at least one of the shown feature-nodes. For every class c with its assigned features $\mathbb{F}^c \in \{1, \dots, n_f^*\}^{n_{\text{wc}}}$, the features are shown in the order of their activation, which can be interpreted as reasoning from the more clearly visible feature like the neck in Figure 2 to the less certain features, such as the pale red eye. The class node is then attached to its last activating feature and the class would be predicted if CHiQPM's calibration or fixed level determines that this feature and all its descendants should be predicted. For a formal definition, we refer the reader to the main paper [8].

2.2.1. Interpretable Conformal Prediction

We introduce interpretable conformal prediction which predicts coherent sets $\mathbb{Y}$ that contain the target label with an error rate $\alpha \in (0, 1)$ while traversing the explaining hierarchy. Towards that goal, we introduce a nonconformity score based on how similar the class is in the hierarchy to the initially predicted class $\hat{c}$, as we are ascending the hierarchy that led to $\hat{c}$. For every class c, we propose to use the activations of the features in the shared path down the tree as nonconformity score:

$$s_{\text{up}}(c) = -\sum_{j=1}^{n_{\text{wc}}} \delta_j^c f^*_{\hat{\mathbb{F}}_j^c} \tag{3}$$

Here, $\delta_j^c \in \{0, 1\}$ encodes if all top j features are shared between c and $\hat{c}$ and $\hat{\mathbb{F}}^c \in \mathbb{F}_c^{n_{\text{wc}}}$ describes the ordered indices of assigned features for every class c by activation in descending order. We call this simple nonconformity score *up* as it goes strictly up the tree.

Table 1. Comparison on Accuracy, Compactness, Contrastiveness and Structural Grounding. Compact describes the number of features n_f^* and features per class n_{wc}. It is binned into very compact + ($n_{\text{wc}} = 5$ and $n_f^* = 50$), medium ∘, and the baseline, denoted - ($n_{\text{wc}} = 2048$ and $n_f^* = 2048$) Among more compact (∘ or above) models, the best result is marked in bold, second best underlined.

Method	Accuracy ↑			Compact	Contrastiveness ↑			SG ↑
	CUB	CAR	IN		CUB	CAR	IN	CUB
Dense Resnet50	86.6	92.1	76.1	-	74.4	75.1	71.6	34.0
glm-saga$_5$	78.0	86.8	58.0	∘	74.0	74.5	71.7	2.5
PIP-Net	82.0	86.5	-	∘	<u>99.5</u>	<u>99.5</u>	-	6.7
ProtoPool	79.4	87.5	-	∘	76.7	78.9	-	13.9
SLDD-Model	84.5	91.1	72.7	+	87.2	89.7	<u>93.4</u>	29.2
Q-SENN	84.7	91.5	<u>74.3</u>	+	93.0	94.2	92.6	23.4
QPM	<u>85.1</u>	<u>91.8</u>	74.2	+	96.0	97.7	89.3	<u>47.9</u>
CHiQPM (**Ours**)	**85.3**	**91.9**	**75.3**	+	**99.9**	**100**	**99.9**	**75.0**

Subtree Selection The conformal predictor can predict with more granularity and achieve its guarantees when the prediction can also go down towards only some subtrees below the feature node determined by s_{up}(c). That also allows CHiQPM to predict those descendants preferably that have some support beyond the shared path, *e.g.* choosing to predict only the cowbirds in Figure 2. Therefore, we extend the nonconformity score to also account for the activation of the feature at the point of diversion after sharing k features:

$$i^{\text{div}} = \hat{\mathbb{F}}^c_{k+1} \quad \text{with} \quad k = \sum_{j=1}^{n_{\text{wc}}-1} \delta_j^c \tag{4}$$

$$s_{\text{sel}}(c) = -f^*_{i^{\text{div}}} - \sum_{j=1}^{n_{\text{wc}}-1} \delta_j^c f^*_{\hat{\mathbb{F}}_j^c} \tag{5}$$

Limited Level Finally, the maximum number of levels the set is constructed from is limited to ensure efficient sets. Towards that goal, the minimum reachable error rate α^n_{cal} on the calibration data for each fixed level n is calculated. The conformal prediction is then limited to the highest level n^{limit} that still reaches the target coverage defined by α. To ensure the limitation, we limit s_{up} to n^{limit}, and multiply the score with the indicator function δ_n^c indexed at $\delta^c_{n^{\text{limit}}}$. This ensures all classes that were not correctly predicted under n^{limit} get the most nonconform nonconformity score of 0[1]:

$$s(c) = \underbrace{\delta^c_{n^{\text{limit}}}}_{\text{Limitation}} \cdot \underbrace{\left(-f^*_{i^{\text{div}}} - \sum_{j=1+n^{\text{limit}}}^{n_{\text{wc}}-1} \delta_j^c f^*_{\hat{\mathbb{F}}_j^c} \right)}_{\text{Limited } s_{\text{sel}}} \tag{6}$$

2.3. Feature Grounding Loss

When two similar classes share $n_{\text{wc}} - 1$ features, Cross-Entropy loss causes only a significant gradient on the single

[1] Equations (4) to (6) constitute a correction of the published work.

Table 2. Average Set Size $|\mathbb{Y}|$ of CHiQPM calibrated to reach various coverages $1-\alpha$ comparing different conformal prediction methods. All methods are very close or reach the desired coverage.

$\lvert\mathbb{Y}\rvert\downarrow$	Inter-	CUB				CARS			INET			
Method	pretable	α=0.12	α=0.1	α=0.075	α=0.05	α=0.075	α=0.05	α=0.0025	α=0.22	α=0.2	α=0.175	α=0.15
Ours	✓	1.22	1.73	2.94	9.05	1.05	1.25	8.25	1.10	1.42	3.25	4.58
$s = s_{\text{sel}}$	✓	4.62	6.15	9.53	29.4	3.62	5.95	28.4	8.14	11.3	17.9	30.5
$s = s_{\text{up}}$	✓	3.03	3.91	8.87	18.7	2.32	3.27	17.9	4.36	6.23	11.8	31.4
THR	✗	1.16	1.32	1.67	2.41	1.02	1.15	2.09	1.05	1.16	1.40	1.87
APS	✗	6.30	7.20	8.54	11.3	5.64	6.83	9.61	16.7	18.9	22.1	26.8

differentiating feature, largely ignoring the shared concepts. To encourage all assigned features to activate meaningfully on the ground truth class while inducing overall sparsity, we propose the Feature Grounding Loss $\mathcal{L}_{\text{feat}}$:

$$\mathcal{L}_{\text{feat}} = -\frac{\sum_{i\in\mathbb{F}} \frac{f_i^*}{|\mathbb{F}|} - \sum_{i\in\overline{\mathbb{F}}} \frac{f_i^*}{|\overline{\mathbb{F}}|}}{\max(f^*)} \tag{7}$$

where $\mathbb{F}$ and $\overline{\mathbb{F}} \in \{1,\dots,n_f^*\}^{n_f^*-n_{\text{wc}}}$ are the indices of the features assigned and not assigned to the ground truth class, respectively. Scaled by the maximum activation, $\mathcal{L}_{\text{feat}}$ balances the gradient across all positively assigned features.

3. Experiments

Following QPM [9], we evaluate our method on the most commonly used datasets for interpretability, CUB-2011 and Stanford Cars [3]. Additionally, ImageNet-1K is used to demonstrate how the method scales to larger problems with more real-world applications. While the proposed method can be applied to any backbone, this extended abstract includes results for Resnet50 [2] with $n_{\text{wc}} = 5$ and $n_f^* = 50$ as main architecture configuration. The main paper [8] includes more extensive results and detailed implementation details. Further, we generally set the density parameter for our class hierarchy to $\rho = 0.5$, as it is sufficient to demonstrate the improvements in built-in set prediction without sacrificing accuracy as point-predictor.

3.1. Metrics

CHiQPM is designed to improve upon QPM, primarily via more easily interpretable class representations, being able to produce meaningful hierarchical explanations and by offering the built-in interpretable calibrated set prediction. Therefore, we evaluate CHiQPM across all QPM metrics in addition to the accuracy as point predictor in relation to its compactness. For evaluating the performance as set predictor, we report the size of the predicted sets, when calibrated to reach a specific accuracy or coverage.

3.2. Results

This section discusses the main quantitative results of our proposed method. Further qualitative examples are included in the conference paper. The accuracy as point predictor along the generally preferable qualities of Compactness, Contrastiveness and Structural Grounding is shown in Tab. 1. CHiQPM shows state-of-the-art accuracy for compact point predictors. Further, it scores nearly perfectly on Contrastiveness. CHiQPM learns features that can be more clearly separated between active and inactive than even the class detectors of *PIP-Net* [6], indicating a gap between the ReLU-induced minimum of 0 and the activations where a relevant concept is found. The clear distinction between active and inactive enables our saliency maps, like in Figures 1 and 2, to also transport *activation* rather than just *location* without a reference test image and therefore enables extensive local explanations in practice. Finally, Structural Grounding quantifies that the additionally added pairs via Equation (2) are also similar in reality and thus lead to more grounded class representations. The state-of-the-art accuracy as point predictor paves the way for accurate set prediction along the hierarchical explanation, as the sets are conditioned on the predicted class.

As comparable CP methods, THR [12] and APS [10] are used, as they are applicable without hyperparameters and broadly used [1]. Table 2 compares our built-in CP method with these and also with the two simpler nonconformity scores s_{sel} and s_{up}. Evidently, our proposed nonconformity score that restricts the sets to be constructed by going up the hierarchical local explanations shows competitive efficiency to THR for higher error rate α and approaches APS for lower values. The reason can be seen when comparing our approach with the simpler s_{sel} that does not restrict the tree level to n^{limit}: With lower α, the gap decreases, as n^{limit} has to be set more loosely, allowing larger and therefore inefficient sets.

4. Conclusion

This work introduces the Calibrated Hierarchical QPM (CHiQPM). Faithfully following its grounded globally inter-

pretable class representations, CHiQPM provides hierarchical local explanations. CHiQPM is calibrated as a form of built-in interpretable Conformal Prediction to traverse the hierarchy at test time and predict a set of coherent classes, similar to how a human reasons, which can be a step towards human-AI complementarity. Finally, CHiQPM's improved global and additional novel form of local interpretability come with state-of-the-art accuracy as compact point predictor and efficiency on par with non-coherent set predictors even on ImageNet-1K, ensuring broad applicability.

Acknowledgements

Financial support for this research was provided by the MWK of Lower Saxony through the Hybrint (VWZN4219) and LCIS (VWZN4704) projects. Furthermore, funding was granted by the Deutsche Forschungsgemeinschaft (DFG) as part of Germany's Excellence Strategy for the PhoenixD (EXC2122) and Quantum Frontiers (EXC2123) Clusters of Excellence, and by the European Union under grant agreement no. 101136006 – XTREME.

References

[1] Alvaro Correia, Fabio Valerio Massoli, Christos Louizos, and Arash Behboodi. An information theoretic perspective on conformal prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 4

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[3] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 4

[4] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956. 1

[5] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019. 1

[6] Meike Nauta, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 4

[7] Thomas Norrenbrock, Marco Rudolph, and Bodo Rosenhahn. Q-senn: Quantized self-explaining neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21482–21491, 2024. 1

[8] Thomas Norrenbrock, Timo Kaiser, Sovan Biswas, Neslihan Kose, Ramesh Manuvinakurike, and Bodo Rosenhahn. CHiQPM: Calibrated hierarchical interpretable image classification. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3, 4

[9] Thomas Norrenbrock, Timo Kaiser, Sovan Biswas, Ramesh Manuvinakurike, and Bodo Rosenhahn. QPM: Discrete optimization for globally interpretable image classification. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 4

[10] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020. 4

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. 2

[12] Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234, 2019. 4

[13] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[14] Eleni Straitouri and Manuel Gomez Rodriguez. Designing decision support systems using counterfactual prediction sets. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2025. 2

[15] Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Improving expert predictions with conformal prediction. In *International Conference on Machine Learning*, pages 32633–32653. PMLR, 2023. 2

[16] Michael Veale and Frederik Zuiderveen Borgesius. Demystifying the draft eu artificial intelligence act—analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4):97–112, 2021. 1

[17] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005. 2

[18] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1