Semantic Consistency in Hierarchical Classification via Probabilistic Logic Constraints

Filippos Gouidis[1,2][a], Antonis Argyros[1,2][b] and Dimitris Plexousakis[1,2][c]
[1]*Computer Science Department,University of Crete, Heraklion, Greece*
[2]*Institute of Computer Science (ICS), Foundation for Research & Technology – Hellas (FORTH), Heraklion, Greece*
{*gouidis,argyros,dp*}*@ics.forth.gr*

Keywords: Hierarchical Classification, Neurosymbolic Learning, Interpretability, Semantic Consistency, Trustworthy AI

Abstract: Deep neural networks trained for hierarchical classification often produce semantically inconsistent predictions, assigning higher confidence to fine-grained classes than to their parent categories. Such violations undermine interpretability and trust, particularly in safety-critical applications. In this work, we propose a lightweight neurosymbolic framework for hierarchical classification that enforces explicit semantic consistency at the probability level during training. Our approach introduces a differentiable logic-based loss that constrains parent-child class probabilities to respect known taxonomic relationships, together with a consensus loss that aligns coarse-grained predictions with aggregated fine-grained evidence. Unlike post-hoc explanation methods, interpretability in our framework emerges intrinsically from the logical structure imposed on the model's output space. Through controlled ablations and cross-backbone evaluations on three benchmark datasets we demonstrate that the proposed semantic-aware training substantially reduces logical violations without degrading classification accuracy. Furthermore, semantic auditing using representational similarity and error depth metrics shows that consistency-enforced models exhibit more rational failure modes, confusing semantically related classes rather than producing implausible errors. These results indicate that simple, explicit semantic constraints can significantly improve the interpretability and trustworthiness of hierarchical classifiers while remaining compatible with standard deep learning architectures.

1 Introduction

The ability to categorize objects at multiple levels of abstraction is a hallmark of human visual intelligence. We recognize an object not just as a distinct entity (e.g., "German Shepherd") but simultaneously as a member of broader semantic groups ("Dog", "Animal"). This hierarchical understanding provides a strong prior that aids in recognizing novel objects; if we know an animal is a bird, we immediately narrow down the search space of possible specific species.

However, standard Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) do not inherently model hierarchical structure in the label space. They are typically trained using a flat Cross-Entropy objective that treats all classes as independent and equidistant (Valmadre, 2022). While effective for closed-set classification, this formulation fails to encode logical entailment relations between parent and child categories. As a result, models frequently exhibit the *inconsistency problem*, assigning high confidence to a fine-grained class while simultaneously assigning low probability to its ancestor (Giunchiglia and Lukasiewicz, 2020). Such violations reflect a lack of semantic coherence rather than insufficient visual discrimination, and are closely related to broader challenges of commonsense reasoning in AI (Marcus, 2020).

Figure 1 illustrates this issue in a simple hierarchical example. A conventional classifier may assign high confidence to a fine-grained category while simultaneously assigning lower confidence to its parent class, violating basic taxonomic structure. Moreover, predictions at different levels of the hierarchy may be internally inconsistent, with aggregated evidence from fine-grained classes failing to align with coarse-grained predictions. Such behavior undermines interpretability and makes model decisions difficult to justify or audit.

This inconsistency is not merely superficial; it is

[a] https://orcid.org/0000-0002-9539-8749
[b] https://orcid.org/0000-0001-8230-3192
[c] https://orcid.org/0000-0002-0863-8266

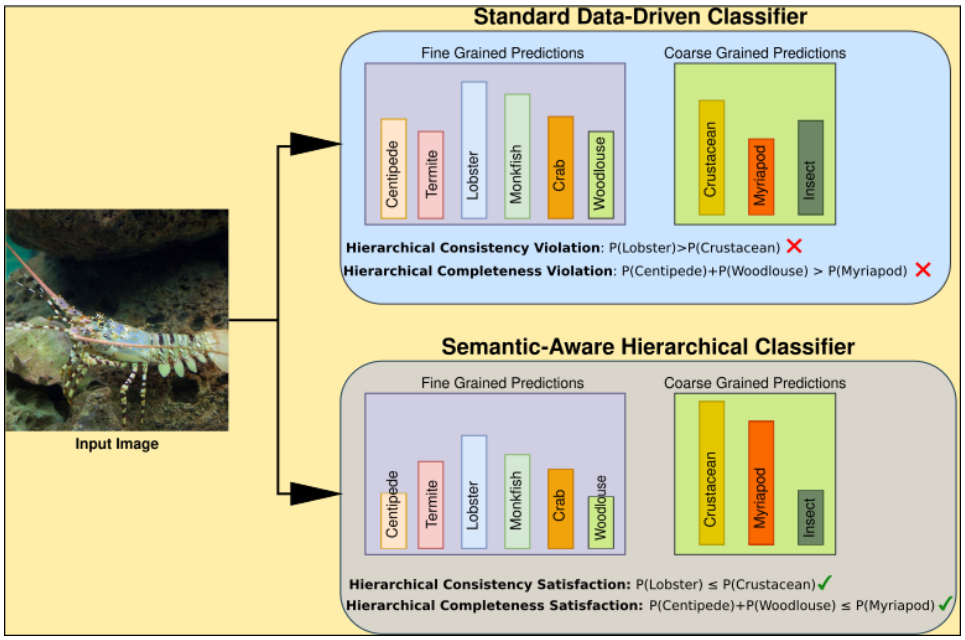

Figure 1: Illustration of Hierarchical Semantic Violations and the NHC Solution. (Top) A standard data-driven classifier produces *hierarchical semantic violations*, including (i) a *hierarchical consistency violation*, where the probability assigned to a fine-grained class (P(Lobster)) exceeds that of its parent class (P(Crustacean)), and (ii) a *hierarchical completeness violation*, where the aggregated probability mass of child classes is not aligned with the corresponding parent prediction. (Bottom) The proposed Neurosymbolic Hierarchical Classification (NHC) framework enforces hierarchical consistency and completeness through explicit probabilistic constraints, resulting in semantically coherent and more interpretable predictions. The observed probability coherence is a consequence of training-time regularization, rather than explicit constraint enforcement at inference.

a symptom of overfitting to local discriminative features (e.g., texture patches) rather than learning holistic semantic concepts and reflects a deeper mismatch between learned probabilities and semantic structure. This becomes particularly evident when models encounter rare or underrepresented classes, where reliance on local discriminative cues leads to semantically implausible predictions. A model that understands the concept of "Vehicle" should be able to roughly categorize a "Forklift" (unseen) as a vehicle, even if it cannot name the specific type.

In this paper, we propose to solve this problem not by using external semantic data (such as Word2Vec or CLIP embeddings), but by **regularizing the internal consistency** of the model's own predictions.

Our experimental evaluation focuses on three core questions central to interpretable and trustworthy hierarchical learning:

1. whether explicit semantic constraints reduce logical inconsistencies,
2. whether such constraints preserve predictive performance across datasets and architectures, and
3. whether improved consistency translates into more semantically rational errors. Accordingly, we evaluate the proposed framework using controlled ablations, cross-backbone generalization, and semantic auditing metrics, rather than exhaustive benchmarking.

Our contributions are threefold:

- We introduce a lightweight neurosymbolic regularization framework for hierarchical classification that enforces explicit probabilistic semantic constraints during training.
- We formulate a differentiable inequality-based logic loss that guarantees parent-child probability consistency at the decision level.
- We demonstrate through controlled ablations and cross-architecture experiments that semantic consistency can be significantly improved without sacrificing accuracy.
- We show that enforcing hierarchical consistency leads to more semantically rational errors, as measured by representational alignment and semantic confusion depth.

Unlike post-hoc explanation techniques (e.g., saliency maps (Ullah et al., 2020) or attribution methods (Jin et al., 2022)), our approach provides *intrinsic interpretability* by enforcing explicit semantic invariants during training. The imposed constraint $P_{\text{child}} \leq$

P_{parent} corresponds directly to a human-understandable "is-a" relationship, ensuring that the model's probabilistic outputs remain semantically valid by construction. This enables straightforward auditing of predictions, where violations of the hierarchy can be explicitly identified, quantified, and traced back to specific decision pathways. As a result, interpretability emerges not from auxiliary explanations, but from the logical structure of the model's output space itself.

2 Related Work

2.1 Hierarchical Classification in Deep Learning

Traditional convolutional neural networks (CNNs) treat output classes as independent and mutually exclusive, ignoring inherent hierarchical relationships among categories. Early works such as (Deng et al., 2011) and (Silla Jr and Freitas, 2011) surveyed hierarchical classification and highlighted the challenges of modeling parent–child dependencies. More recently, (Wu et al., 2019) analyzed the issues of applying hierarchical loss functions when classes are not strictly tree-structured, emphasizing the need for semantic consistency. Weber et al. (Weber et al., 2024) addressed hierarchical bias in semantic segmentation by embedding class hierarchies in hyperbolic space, demonstrating improved structure preservation. However, these methods primarily reformulate embeddings or hierarchical metrics, rather than enforcing direct probabilistic constraints between parent and child classes. Additionally, hierarchical softmax and structured loss formulations have been proposed to exploit taxonomic structure during training (Valmadre, 2022). While effective for large-scale classification, these approaches primarily restructure the output space or optimization procedure, and do not explicitly enforce probabilistic semantic invariants between parent and child classes. Goren et al. (Goren et al., 2024) studied hierarchical selective classification, allowing models to abstain at fine-grained levels under uncertainty, whereas our work focuses on enforcing semantic consistency among predicted probabilities during training. Kwon et al. (Kwon et al., 2024) modeled visual hierarchies using hyperbolic embeddings, capturing hierarchical structure implicitly in feature space. Our approach differs by enforcing semantic consistency explicitly at the probability level, independent of the embedding geometry.

2.2 Semantic Consistency and Logical Constraints

Several studies have explored incorporating symbolic or logical knowledge into neural models. Xu et al. (Xu et al., 2018) introduced the *Semantic Loss*, which integrates symbolic constraints into deep learning objectives, showing that logical consistency can guide learning toward semantically valid outcomes. Similarly, (Fischer et al., 2019) proposed DL2, a framework that enforces differentiable logical constraints in neural networks. These methods inspired subsequent neurosymbolic models that bridge symbolic reasoning and statistical learning. In the context of hierarchical classification, recent work by (Wu et al., 2025) introduced dependency-based constraints to reduce inconsistency in fine-grained recognition. Kull et al. (Kull et al., 2019) show that standard post-hoc calibration methods can improve probability estimates but do not address structural or semantic incoherence in multi-class predictions. However, most existing approaches enforce semantic structure indirectly, either through architectural priors, custom embeddings, or global logical formulations, rather than by explicitly constraining probability mass between hierarchy levels. In contrast, our Semantic Consistency Loss (L_{Logic}) directly penalizes local parent–child violations by enforcing simple inequality constraints on predicted probabilities. This formulation is lightweight, scalable to modern deep vision models, and enables fine-grained auditing of semantic inconsistencies at the level of individual predictions.

2.3 Neurosymbolic and Knowledge-Guided Reasoning in Vision

Neurosymbolic learning combines symbolic reasoning with neural models to improve interpretability and semantic validity. Foundational works by (Garcez et al., 2019) and (Besold et al., 2017) showed how logical constraints can guide learning beyond purely data-driven correlations.

In computer vision, many recent works have explored the use of structured knowledge across a variety of vision-based tasks (Gouidis et al., 2020). More closely related to the scope of this paper, such approaches have been applied to hierarchical label spaces, where semantic entailment relations such as *is-a* play a central role. Shen et al. (Shen et al., 2019) demonstrated improved interpretability through hierarchical semantic CNNs in medical imaging. Bertinetto et al. (Bertinetto et al., 2020) introduced the notion of

making "better mistakes," encouraging errors to remain within semantically related subtrees. Giunchiglia et al. (Giunchiglia and Lukasiewicz, 2020) enforced hard hierarchical constraints to guarantee consistency in multi-label prediction, while (Park et al., 2024) extended hierarchical reasoning to dense prediction tasks. Our approach enforces semantic coherence via soft, differentiable constraints on output probabilities, avoiding architectural changes or hard logic while remaining compatible with standard CNNs and Vision Transformers.

3 Methodology

Our approach integrates symbolic knowledge concerning class hierarchy directly into the optimization process of a deep neural network. Overall, our method is based on the following components.

Dual-Head Architecture: We employ a versatile feature extractor Φ, supporting diverse backbones including standard CNNs (ResNet), parameter-efficient models (EfficientNet), and attention-based Vision Transformers (Swin, ViT). The resulting feature vector $f \in \mathbb{R}^D$ is processed by two independent classification heads, H_{Fine} and H_{Coarse}, to produce fine- and coarse-grained logits. We assume a two-level hierarchy in which each fine-grained class is associated with a single coarse parent.

Supervised Learning Objectives: Both classification heads are supervised using standard Cross-Entropy (CE) loss. Given ground-truth labels y_{fine} and y_{coarse}, the supervised losses are defined as:

$$\mathcal{L}_{CE}^{fine} = -\frac{1}{B}\sum_{b=1}^{B} \log(P_{fine}^{(b)}(y_{fine}^{(b)})) \tag{1}$$

$$\mathcal{L}_{CE}^{coarse} = -\frac{1}{B}\sum_{b=1}^{B} \log(P_{coarse}^{(b)}(y_{coarse}^{(b)})) \tag{2}$$

Here, P_{fine} and P_{coarse} denote the Softmax-normalized probability distributions derived from the corresponding logits.

Consistency Regularization:

- **Semantic Consistency Loss:** The logical constraint is based on the "Is-A" hierarchy: the probability of a child class i must not exceed the probability of its parent $j(i)$. We enforce this by penalizing only positive differences (violations).

$$\mathcal{L}_{Logic} = \frac{1}{B \cdot N_{fine}} \sum_{b=1}^{B}\sum_{i=1}^{N_{fine}} \mathcal{V}_{b,i}, \tag{3}$$

$$\mathcal{V}_{b,i} = \mathrm{ReLU}(P_{fine}^{(b)}(i) - P_{coarse}^{(b)}(j(i))) \tag{4}$$

We enforce semantic consistency directly on *Softmax probabilities* rather than logits or latent embeddings. This ensures that logical constraints operate at the decision level, making the final outputs interpretable as calibrated semantic beliefs that respect the underlying taxonomy.

- **Consensus Loss:** While $\mathcal{L}_{Logic}$ enforces an inequality, $\mathcal{L}_{Consensus}$ ensures probabilistic completeness by encouraging coarse predictions to match the sum of their constituent fine-grained probabilities. We formalize this using a binary hierarchy matrix $M \in \{0,1\}^{N_{fine} \times N_{coarse}}$, where $M_{i,j} = 1$ if fine class i is a child of superclass j, and 0 otherwise. The derived coarse distribution is calculated as:

$$P_{derived} = P_{fine} \cdot M \tag{5}$$

The Consensus Loss is then defined as the squared error between the predicted and derived distributions:

$$\mathcal{L}_{Consensus} = ||P_{coarse} - P_{derived}||_2^2 \tag{6}$$

where N_{fine} and N_{coarse} correspond to the number of fine-grained and coarse-grained classes respectively. The probabilities for the predicted classes are derived via Softmax operation. We employ a mean squared error objective for $L_{\text{Consensus}}$ to promote stable alignment between predicted and derived coarse distributions. Compared to divergence-based alternatives (e.g., KL divergence), this symmetric formulation avoids instability when probabilities approach zero and empirically leads to smoother optimization in multi-head architectures.

Total Loss Function The final training objective combines statistical performance with symbolic constraints:

$$\mathcal{L}_{Total} = \mathcal{L}_{CE}^{fine} + \mathcal{L}_{CE}^{coarse} + \lambda\mathcal{L}_{Logic} + \gamma\mathcal{L}_{Consensus} \tag{7}$$

where λ and γ control the regularization strength.

4 Experimental Evaluation

4.1 Evaluation Dimensions and Metrics

To move beyond flat classification accuracy, we evaluate our models through three complementary "Dimensions of Trust." This taxonomy allows us to audit the internal structural integrity of the learned manifold rather than just its predictive output:

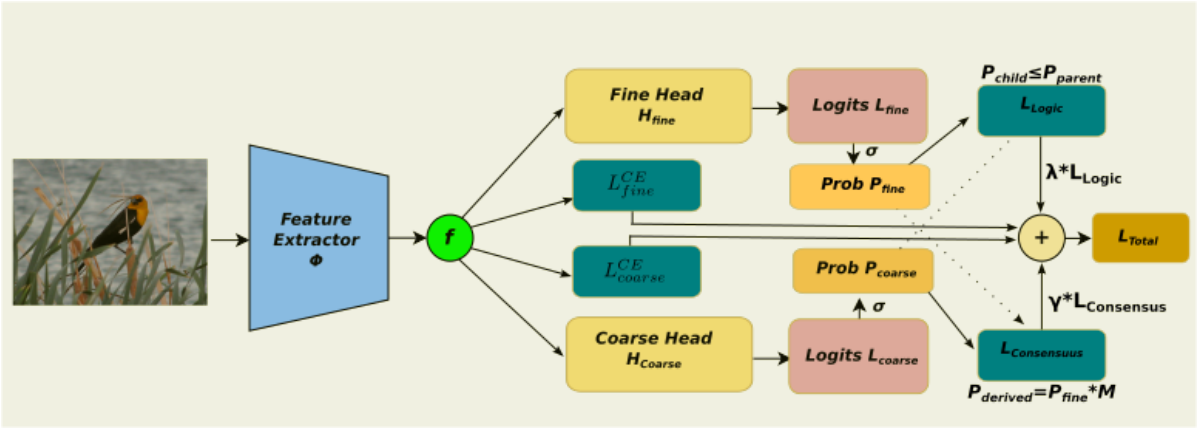

Figure 2: Overview of the proposed Neurosymbolic Hierarchical Classification (NHC) framework. A shared feature extractor feeds two classification heads for fine-grained and coarse-grained prediction. Standard cross-entropy losses supervise both heads, while two symbolic regularizers enforce semantic consistency: (i) a logic-based constraint ensuring that child class probabilities do not exceed those of their parent classes, and (ii) a consensus loss aligning coarse predictions with aggregated fine-grained evidence. All components are optimized jointly via the total loss.

– **Generalization Robustness**: Evaluates standard classification performance across datasets and architectures to assess whether semantic constraints generalize beyond a single model configuration.
– **Logical Consistency:** Assesses the model's adherence to the fundamental "is-a" constraints of the symbolic hierarchy, quantifying the frequency and severity of predictions that violate taxonomic rules.
– **Semantic Rationality:** Evaluates the geometric and biological alignment of the model's internal feature space with human-interpretable structures, ensuring that error modes are semantically plausible.

This categorization reflects the interpretability and trust-oriented goals of the proposed framework. The different metrics are defined as follows.

1. Classification Accuracy: We report the standard Top-1 accuracy for both fine-grained (Acc_{Fine}) and coarse-grained (Acc_{Coarse}) predictions.

2. i) Logical Violation Rate: We formally define the **Logical Violation Rate** ($\mathcal{LVR}$) based on the model's test set predictions. A logical violation occurs when the model predicts a specific fine class i and assigns a higher probability to it than to its true parent coarse class $j(i)$.

A single prediction x is considered a logical violation if:

$$P_{fine}(\hat{i}|x) > P_{coarse}(j(\hat{i})|x) \tag{8}$$

where $\hat{i}$ is the index of the predicted fine class (argmax). The total $\mathcal{LVR}$ over the entire test set $\mathcal{D}$ is calculated as:

$$\mathcal{LVR} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbb{I}\left[P_{fine}(\hat{i}|x) > P_{coarse}(j(\hat{i})|x)\right] \tag{9}$$

where $\mathbb{I}[\cdot]$ is the indicator function.

2. ii) Average Penalty Magnitude: Additionally, we track the **Average Penalty Magnitude** ($Avg(\mathcal{L}_{Logic})$), which quantifies the *severity* of these violations. It is measured as the mean positive difference between child and parent probabilities across the test set:

$$Avg(\mathcal{L}_{Logic}) = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \text{ReLU}(P_{fine}(\hat{i}) - P_{coarse}(j(\hat{i}))) \tag{10}$$

$$MSE_{Comp} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} ||\mathbf{p}_{coarse} - (\mathbf{p}_{fine} \cdot \mathcal{M})||_2^2 \tag{11}$$

where $\mathcal{M}$ is the fixed binary hierarchy matrix. This metric quantifies the "Semantic Gap" between the two heads; a high MSE_{Comp} indicates that the coarse-grained head is making decisions that are numerically unsupported by the fine-grained features.

3.i) RSA Alignment: To quantify the interpretability of the learned latent space, we employ **Representational Similarity Analysis (RSA)**, following the analysis by Kornblith et in (Kornblith et al., 2019), who demonstrated that pairwise distance correlations provide a robust and architecture-agnostic measure of representation alignment. We compute the correlation between

Table 1: Ablation study on CIFAR-100 (ResNet-50). Evaluation of the effect of the semantic logic (λ) and the consensus (γ) losses, individually and jointly. **Bold** and underlined values denote the best and second-best performance, respectively.

Regularization Terms		**Hierarchy Consistency Metrics**			**Accuracy**	
λ	γ	**LVR** $\downarrow$	**APM** $\downarrow$	**MSE** $\downarrow$	$\mathbf{Acc}_{Fine}$ $\uparrow$	$\mathbf{Acc}_{Coarse}$ $\uparrow$
Baseline Models (No Consistency)						
0.0	0.0	28.3	10.1	$2.5e-3$	77.5 ± 1.2	77.9 ± 1.3
Ablation for Semantic Consistency (λ Term)						
0.01		28.1	9.2	$3.0e-3$	75.5 ± 0.4	75.9 ± 0.4
0.1	0.0	26.3	7.3	$2.8e-3$	75.6 ± 0.2	76.4 ± 0.3
1.0		**22.1**	7.0	$2.6e-3$	76.7 ± 0.3	77.2 ± 0.3
Ablation for Probabilistic Completeness (γ Term)						
	0.01	42.1	10.4	$3.8e-3$	76.1 ± 0.3	76.3 ± 0.3
0.0	0.1	34.5	8.5	$3.0e-3$	76.6 ± 0.5	76.7 ± 0.3
	1.0	24.5	**6.2**	$\mathbf{1.6e-3}$	$\mathbf{77.7 \pm 0.4}$	77.9 ± 0.6
Ablation for Both Terms						
0.01	0.01	30.2	7.3	$2.7e-3$	75.8 ± 0.4	76.3 ± 0.7
0.01	0.1	26.1	7.2	$2.3e-3$	77.2 ± 0.4	77.9 ± 0.7
0.01	1.0	24.8	6.6	$1.9e-3$	76.7 ± 0.4	77.7 ± 0.7
0.1	0.01	26.6	7.9	$2.5e-3$	75.4 ± 0.4	75.7 ± 0.6
0.1	0.1	24.6	7.5	$2.4e-3$	76.5 ± 0.4	77.4 ± 0.7
0.1	1.0	25.1	7.3	$2.1e-3$	77.3 ± 0.4	$\mathbf{78.0 \pm 0.7}$
1.0	0.01	28.1	7.5	$2.7e-3$	76.4 ± 0.4	76.8 ± 0.7
1.0	0.1	28.6	8.5	$2.8e-3$	76.5 ± 0.4	76.7 ± 0.7
1.0	1.0	30.1	7.9	$2.3e-3$	75.7 ± 0.4	76.7 ± 0.7

the pairwise distances of fine-grained class prototypes in the learned feature space and the corresponding ground-truth taxonomic distances D_C. . This measures the degree to which the model's internal "mental map" aligns with human logic. Let $\mathbf{D}_{model}$ and $\mathbf{D}_{GT}$ be the pairwise distance matrices for the model archetypes and the symbolic ground truth, respectively. The alignment is defined as:

$$\rho_{RSA} = \text{Spearman}\left(\text{vec}(\mathbf{D}_{model}), \text{vec}(\mathbf{D}_{GT})\right) \quad (12)$$

where $\text{vec}(\cdot)$ extracts the unique upper-triangular elements of the matrices. A score near 1.0 indicates perfect structural alignment, while a score near 0 suggests that the model's internal feature space is semantically incoherent with respect to the known taxonomy.

3.ii) Semantic Confusion Index (SCI): To measure the "rationality" of misclassifications, we define the **SCI** as the average taxonomic distance between the predicted label and the true label for all incorrect predictions. For a hierarchy with taxonomic distance matrix D_C, the SCI is calculated as:

$$SCI = \frac{1}{|\mathcal{D}_{err}|} \sum_{(x,y)\in\mathcal{D}_{err}} D(y, \hat{y}) \quad (13)$$

where D_{err} denotes the set of samples for which the predicted fine-grained label $\hat{y}$ does not match the ground truth y. A lower SCI indicates that the model's errors are "closer" in the taxonomic tree (e.g., confusing two bird species within the same family).

We emphasize that our evaluation prioritizes interpretability, consistency, and semantic behavior over marginal gains in top-1 accuracy. As such, we do not aim to exhaustively compare against all hierarchical classification baselines, but instead focus on isolating the effect of explicit semantic constraints within standard deep learning architectures.

4.2 Experimental Setup

Datasets: We evaluate our approach on three widely used benchmarks with hierarchical annotations:

Table 2: Ablation study on CUB-200-2011 (ResNet-50). Evaluation of the effect of the semantic logic (λ) and the consensus (γ) losses, individually and jointly. **Bold** and underlined values denote the best and second-best performance, respectively.

Regularization Terms		**Hierarchy Consistency Metrics**			**Accuracy**	
$\boldsymbol{\lambda}$	$\boldsymbol{\gamma}$	**LVR** $\downarrow$	**APM** $\downarrow$	**MSE** $\downarrow$	$\textbf{Acc}_{Fine}$ $\uparrow$	$\textbf{Acc}_{Coarse}$ $\uparrow$
Baseline Models (No Consistency)						
0.0	0.0	46.8	28.4	$16.6e-3$	35.4 ± 0.2	60.4 ± 1.3
Ablation for Semantic Consistency (λ Term)						
0.01		41.4	26.2	$13.0e-3$	33.8 ± 0.3	61.0 ± 0.4
0.1	0.0	39.4	22.9	$16.5e-3$	38.6 ± 0.2	62.8 ± 0.3
1.0		**31.4**	27.8	$12.5e-3$	33.0 ± 0.3	60.9 ± 0.3
Ablation for Probabilistic Completeness (γ Term)						
	0.01	43.7	27.0	$12.8e-3$	35.8 ± 0.2	60.8 ± 0.3
0.0	0.1	39.9	23.8	$11.0e-3$	$\mathbf{38.9 \pm 0.2}$	$\underline{63.1 \pm 0.2}$
	1.0	38.6	$\underline{22.2}$	$\mathbf{9.1e-3}$	35.4 ± 0.3	61.2 ± 0.4
Ablation for Both Terms						
0.01	0.01	40.0	23.1	$15.4e-3$	35.6 ± 0.2	62.5 ± 0.3
0.01	0.1	47.2	28.3	$16.5e-3$	35.4 ± 0.4	56.7 ± 0.4
0.01	1.0	40.4	26.2	$\underline{10.7e-3}$	35.0 ± 0.3	60.4 ± 0.5
0.1	0.01	45.0	27.3	$14.2e-3$	36.0 ± 0.4	59.4 ± 0.6
0.1	0.1	$\underline{35.3}$	**20.0**	$14.5e-3$	$\underline{38.8 \pm 0.4}$	$\mathbf{63.3 \pm 0.5}$
0.1	1.0	45.4	30.3	$13.7e-3$	33.4 ± 0.3	54.4 ± 0.5
1.0	0.01	41.9	29.8	$11.2e-3$	33.1 ± 0.4	59.6 ± 0.5
1.0	0.1	41.7	29.7	$11.1e-3$	32.6 ± 0.3	58.9 ± 0.4
1.0	1.0	39.6	26.1	$11.6e-3$	30.9 ± 0.3	58.8 ± 0.4

– **CIFAR-100** (Krizhevsky and Hinton, 2009) consists of 100 fine-grained object classes grouped into 20 superclasses and serves as a controlled benchmark for hierarchical consistency.
– **CUB-200-2011** (Wah et al., 2011) is a fine-grained bird recognition dataset with 200 species organized into higher-level taxonomic groups, posing a challenging setting with subtle inter-class variations.
– **AwA2** (Xian et al., 2017) contains 50 animal categories grouped into 10 coarse classes and is widely used to study semantic structure and attribute-based generalization. AwA2 is typically evaluated in a zero-shot setting with disjoint seen and unseen classes. Here, we adopt the standard supervised protocol and train and test exclusively on the seen classes, as our goal is to analyze semantic consistency within a fixed hierarchical taxonomy rather than zero-shot transfer.

An overview of the datasets and their key characteristics is presented in Table 6.

Backbone Architectures: We evaluate our framework across a diverse set of backbones ranging from standard CNNs to modern Vision Transformers:

– **Standard CNNs:** We employ **ResNet-18** and **ResNet-50** (He et al., 2016) as representative deep residual networks.
– **Efficient CNNs:** We utilize **EfficientNet-B0** and **EfficientNet-B1** (Tan and Le, 2019) to test performance under parameter-constrained settings.
– **Vision Transformers:** We evaluate on **ViT-B/16** (Dosovitskiy et al., 2021) and **Swin-T** (Liu et al., 2021) to assess if logic constraints are effective on patch-based, attention-driven architectures.

Training Details: All models were trained using backbone- and dataset-specific hyperparameters to ensure stable convergence and fair comparison. Unless otherwise stated, inputs were resized to 224×224, which is standard for ImageNet-pretrained architectures. For CIFAR-100, ResNet-18 was additionally evaluated using the native 32×32 resolution as a traditional baseline, while

Table 3: Ablation study on AwA2 (ResNet-50). Evaluation of the effect of the semantic logic (λ) and the consensus (γ) losses, individually and jointly. **Bold** and underlined values denote the best and second-best performance, respectively.

Regularization Terms		Hierarchy Consistency Metrics			Accuracy	
λ	γ	LVR ↓	APM ↓	MSE ↓	Acc_{Fine} ↑	Acc_{Coarse} ↑
Baseline Models (No Consistency)						
0.0	0.0	55.6	4.2	$2.0e-3$	87.8 ± 0.9	88.6 ± 1.1
Ablation for Semantic Consistency (λ Term)						
0.01		46.2	4.4	$1.5e-3$	87.4 ± 0.6	88.7 ± 0.7
0.1	0.0	42.3	4.1	$1.9e-3$	87.3 ± 0.8	89.3 ± 0.8
1.0		41.0	4.4	$1.7e-3$	$\mathbf{88.1 \pm 0.6}$	89.3 ± 0.6
Ablation for Probabilistic Completeness (γ Term)						
	0.01	48.9	4.3	$1.2e-3$	87.9 ± 0.8	$\mathbf{89.9 \pm 0.8}$
0.0	0.1	54.7	3.9	$1.1e-3$	87.6 ± 0.2	89.6 ± 0.2
	1.0	46.2	**3.4**	$1.1e-3$	87.8 ± 0.3	89.4 ± 0.4
Ablation for Both Terms						
0.01	0.01	54.5	4.2	$1.7e-3$	88.0 ± 0.7	89.1 ± 0.8
0.01	0.1	47.8	4.5	$1.6e-3$	87.5 ± 0.8	88.8 ± 0.7
0.01	1.0	42.3	**3.4**	$\mathbf{0.9e-3}$	87.3 ± 0.6	89.0 ± 0.7
0.1	0.01	52.3	5.6	$1.5e-3$	86.3 ± 0.7	87.2 ± 0.6
0.1	0.1	52.8	3.9	$1.6e-3$	85.9 ± 0.4	88.5 ± 0.5
0.1	1.0	**38.0**	4.4	$1.2e-3$	87.0 ± 0.8	88.4 ± 07
1.0	0.01	50.6	5.6	$1.5e-3$	86.7 ± 0.7	88.3 ± 0.8
1.0	0.1	41.9	5.3	$1.2e-3$	87.7 ± 0.7	89.7 ± 0.7
1.0	1.0	44.5	3.8	$1.0e-3$	87.3 ± 0.8	89.5 ± 0.7

all other backbones operated on resized inputs.

Convolutional networks were optimized using stochastic gradient descent (SGD) with an initial learning rate of 0.01, except for ResNet-18 on CIFAR-100, which used a learning rate of 0.1 following common practice. Fine-grained experiments on CUB-200-2011 employed the Adam optimizer to facilitate stable fine-tuning. Transformer-based models were optimized using AdamW with a learning rate of 3×10^{-5}.

4.3 Results Overview

4.3.1 Logic Violation Reduction

The results on CIFAR-100 using a ResNet-50 backbone are presented in Table 1. Without any hierarchical regularization, the baseline model exhibits a high $\mathcal{LVR}$ of 28.3%, confirming that flat training frequently produces semantically invalid predictions. Introducing the logic-based constraint alone ($\lambda > 0, \gamma = 0$) yields a monotonic reduction in violation severity ($Avg(\mathcal{L}_{Logic})$) and frequency, but does not fully eliminate inconsistencies. In contrast, jointly optimizing L_{Logic} and $L_{\text{Consensus}}$ leads to a substantial drop in $\mathcal{LVR}$ (up to ∼15-20% relative reduction), while preserving fine-grained accuracy. This demonstrates that inequality-based constraints are most effective when supported by probabilistic completeness between hierarchy levels.

The effect is even more pronounced on the fine-grained CUB-200-2011 dataset (Table 2 and Table 4). As the baseline model violates hierarchical constraints in nearly half of all predictions ($\mathcal{LVR}$ = 46.8%), introducing semantic consistency losses reduces violations by a large margin in the optimal regime (e.g., $\lambda = 0.1, \gamma = 1.0$), without degrading coarse-grained accuracy.

Notably, the $Avg(\mathcal{L}_{Logic})$ and MSE_{Comp} metrics decrease in tandem, indicating that the coarse head becomes numerically grounded in fine-grained evidence rather than acting as an independent classifier. This confirms that the proposed losses act as structural regularizers rather than post-hoc corrections.

Similar trends are observed on AwA2 (Table 3). Baseline models already achieve high fine- and

Table 4: Cross-backbone evaluation on CIFAR-100, CUB-200-2011 and Awa2. Baseline models are trained without semantic constraints ($\lambda = 0, \gamma = 0$), while NHC enforces hierarchical consistency ($\lambda = 0.1, \gamma = 1$). **Bold** values denote the best result per backbone.

Dataset	Backbone	Type	Hierarchy Consistency Metrics			Accuracy	
			LVR ↓	APM ↓	MSE ↓	Acc_{Fine} ↑	Acc_{Coarse} ↑
CIFAR-100	ResNet-18	Baseline	25.6	9.9	$24.3e-4$	60.5 ± 0.3	62.8 ± 0.4
		NHC	**22.9**	**9.5**	$\mathbf{2.1e-4}$	$\mathbf{61.3 \pm 0.2}$	$\mathbf{63.7 \pm 0.3}$
	EfficientNet-B0	Baseline	49.5	9.1	$3.3e-4$	77.1 ± 0.4	77.6 ± 0.5
		NHC	**43.5**	**8.1**	$\mathbf{2.9e-4}$	$\mathbf{79.1 \pm 0.4}$	$\mathbf{79.7 \pm 0.6}$
	EfficientNet-B1	Baseline	43.2	7.1	$2.8e-4$	80.6 ± 0.4	81.2 ± 0.5
		NHC	**37.4**	**6.2**	$\mathbf{2.3e-4}$	$\mathbf{81.2 \pm 0.4}$	$\mathbf{85.5 \pm 0.4}$
	Swin-T	Baseline	42.0	26.1	$12.2e-4$	33.4 ± 0.4	59.2 ± 0.6
		NHC	**38.4**	**23.2**	$\mathbf{11.5e-4}$	$\mathbf{35.8 \pm 1.2}$	$\mathbf{61.0 \pm 1.3}$
	ViT-B/16	Baseline	45.3	23.6	$16.6e-4$	49.6 ± 0.4	66.0 ± 0.3
		NHC	**35.0**	**20.0**	$\mathbf{14.8e-4}$	$\mathbf{43.8 \pm 0.4}$	$\mathbf{66.3 \pm 0.4}$
CUB-200	ResNet-18	Baseline	46.8	34.4	$15.7e-4$	31.6 ± 0.5	56.8 ± 1.3
		NHC	**34.1**	**24.5**	$\mathbf{10.4e-4}$	$\mathbf{36.4 \pm 0.2}$	$\mathbf{59.3 \pm 0.4}$
	EfficientNet-B0	Baseline	42.1	26.2	$11.8e-4$	33.4 ± 0.5	59.2 ± 0.6
		NHC	**37.3**	**21.6**	$\mathbf{10.4e-4}$	$\mathbf{37.2 \pm 0.5}$	$\mathbf{62.3 \pm 0.5}$
	EfficientNet-B1	Baseline	46.8	28.4	$16.6e-4$	$\mathbf{36.4 \pm 0.7}$	60.4 ± 0.6
		NHC	**35.3**	**26.3**	$\mathbf{10.7e-4}$	35.8 ± 1.2	60.4 ± 1.3
	Swin-T	Baseline	42.5	11.3	$11.3e-4$	$\mathbf{56.4 \pm 0.4}$	59.2 ± 0.6
		NHC	**28.4**	**33.2**	$\mathbf{6.5e-4}$	55.8 ± 0.9	$\mathbf{61.0 \pm 1.0}$
	ViT-B/16	Baseline	45.3	23.6	$16.6e-4$	49.6 ± 0.4	66.0 ± 0.3
		NHC	**28.2**	**13.6**	$\mathbf{8.3e-4}$	$\mathbf{57.8 \pm 0.4}$	$\mathbf{74.5 \pm 0.4}$
AwA2	ResNet-18	Baseline	50.0	4.9	$1.9e-3$	86.3 ± 0.6	88.2 ± 0.7
		NHC	**43.3**	**3.8**	$\mathbf{1.3e-3}$	$\mathbf{86.7 \pm 0.7}$	$\mathbf{88.5 \pm 0.6}$
	EfficientNet-B0	Baseline	53.4	4.5	$1.1e-3$	91.7 ± 0.4	$\mathbf{92.9 \pm 0.6}$
		NHC	**42.1**	**4.0**	$\mathbf{1.0e-3}$	$\mathbf{92.1 \pm 0.5}$	92.5 ± 0.5
	EfficientNet-B1	Baseline	44.5	7.6	$1.0e-3$	$\mathbf{93.2 \pm 0.7}$	93.6 ± 0.6
		NHC	**35.6**	**5.7**	$\mathbf{0.5e-3}$	93.2 ± 0.6	$\mathbf{94.2 \pm 0.6}$
	Swin-T	Baseline	34.5	4.7	$0.5e-4$	$\mathbf{95.2 \pm 0.4}$	$\mathbf{96.0 \pm 0.6}$
		NHC	**29.6**	**3.7**	$\mathbf{0.4e-3}$	95.1 ± 0.4	95.9 ± 0.3
	ViT-B/16	Baseline	32.9	2.9	$0.4e-3$	93.5 ± 0.5	94.6 ± 0.6
		NHC	**29.3**	**2.3**	$\mathbf{0.2e-3}$	$\mathbf{93.6 \pm 0.5}$	$\mathbf{94.9 \pm 0.5}$

Table 5: Semantic rationality audit based on Representational Similarity Analysis (RSA) and the Semantic Confusion Index (SCI). Results compare Baseline models trained without hierarchical constraints ($\lambda = 0, \gamma = 0$) against semantic-aware models trained with NHC ($\lambda = 0.1, \gamma = 1$). **Bold** values denote the best performance for each model.

Dataset	Backbone	RSA (Baseline)↑	RSA (NHC)↑	SCI (Baseline)↓	SCI (NHC)↓
CIFAR-100	**ResNet-18**	0.123	**0.232**	2.16	**1.92**
	ResNet-50	0.331	**0.373**	1.96	**1.92**
	EfficientNet-B0	0.235	**0.265**	1.95	**1.93**
	EfficientNet-B1	0.247	**0.268**	1.96	**1.92**
	Swin-T	0.226	**0.251**	1.92	**1.85**
	ViT-B/16	0.248	**0.256**	1.95	**1.91**
CUB-200	**ResNet-18**	0.071	**0.124**	1.07	**1.05**
	ResNet-50	0.092	**0.105**	1.09	**1.06**
	EfficientNet-B0	0.082	**0.154**	1.02	**0.98**
	EfficientNet-B1	0.099	**0.155**	1.00	**0.93**
	Swin-T	0.048	**0.069**	**0.89**	**0.89**
	ViT-B/16	0.035	**0.052**	1.27	**1.25**
AwA2	**ResNet-18**	0.418	**0.455**	1.63	**1.60**
	ResNet-50	0.481	**0.502**	1.82	**1.67**
	EfficientNet-B0	**0.481**	0.477	1.75	**1.65**
	EfficientNet-B1	0.432	**0.441**	1.68	**1.72**
	Swin-T	0.499	**0.506**	1.75	**1.72**
	ViT-B/16	0.389	**0.487**	**1.67**	**1.67**

coarse-grained accuracy, reflecting the lower visual ambiguity and the shallower hierarchical structure of the dataset. However, despite this strong predictive performance, unconstrained models still exhibit a high rate of logical violations, indicating that accuracy alone does not guarantee semantic coherence.

Introducing semantic consistency constraints consistently reduces the Logical Violation Rate across all configurations, with the strongest improvements obtained when jointly optimizing the logic and consensus losses. Importantly, these gains are achieved without degrading classification accuracy, which remains stable across all settings. This demonstrates that, even in comparatively simple taxonomies, hierarchical regularization improves the structural validity of model predictions rather than compensating for classification difficulty.

4.3.2 Backbone-Agnostic Consistency and Robustness

Table 4 evaluate whether the proposed semantic constraints generalize across different network architectures, including convolutional models (ResNet, EfficientNet) and attention-based Transformers (ViT, Swin). For each backbone, we compare standard training without hierarchical constraints ($\lambda = 0, \gamma = 0$) against our semantic-aware formulation (NHC, $\lambda = 0.1, \gamma = 1$).

Across all architectures and datasets, NHC consistently reduces the $\mathcal{LVR}$ and $Avg(\mathcal{L}_{Logic})$, demonstrating that the proposed constraints are not tied to a specific inductive bias. Notably, Transformer-based models exhibit the largest relative improvements in consistency. While these models achieve competitive accuracy under standard training, they produce highly inconsistent probability assignments, likely due to the absence of an explicit hierarchical inductive bias. Introducing semantic constraints compensates for this limitation by explicitly enforcing taxonomic structure at the output level.

Importantly, improvements in logical consistency do not come at the cost of predictive performance. Across all backbones, fine- and coarse-grained accuracies are preserved or slightly improved under semantic-aware training. This indicates that the proposed constraints act as regularizers that stabilize optimization rather than restricting model expressiveness. From an interpretability perspective, this backbone-agnostic behavior is crucial: it shows that semantic validity can be enforced independently of architectural choices, making the approach applicable to a wide range of practical systems.

Table 6: Summary of datasets used in the experimental setting. Image counts refer to the full dataset; CIFAR-100 and CUB-200-2011 follow standard train/test splits. For AwA2, train/test splits refer only to the seen classes set.

Dataset	Images	N_{fine}	N_{coarse}	Hierarchy Levels	Domain
CIFAR-100	60,000	100	20	2	Generic objects
CUB-200-2011	11,788	200	30	2	Fine-grained birds
AwA2	29,800*	40*	5*	2	Animals

4.3.3 Semantic Rationality and Error Depth

The analysis of semantic rationality is summarized in Table 5. RSA measures the alignment between the geometry of the learned representation space and the ground-truth taxonomic structure, while SCI quantifies the semantic severity of misclassifications.

Across all datasets and backbones, NHC consistently improves RSA, indicating that enforcing hierarchical consistency at the output level reshapes the internal feature space toward human-interpretable structure. This effect is most pronounced on CUB-200-2011, where fine-grained distinctions induce a complex taxonomic geometry that is poorly captured by unconstrained models.

SCI results further reveal dataset-dependent behavior. On CIFAR-100 and CUB-200-2011, semantic-aware models consistently reduce SCI, demonstrating that errors remain closer to the correct superclass or family. This reflects graceful degradation: even when predictions are incorrect, they respect semantic proximity.

On AwA2, absolute SCI values are lower due to the shallower hierarchy and fewer coarse categories. Nevertheless, NHC still yields systematic reductions or stabilization of SCI across most backbones, indicating improved rationality of failure modes even in settings where semantic distances are compressed. Importantly, SCI values are not intended for cross-dataset comparison, but for within-dataset behavioral auditing.

Taken together, improvements in RSA and SCI provide complementary evidence that hierarchical consistency constraints enhance not only probabilistic validity, but also the qualitative structure of learned representations and errors.

5 Conclusion

In this work we presented a lightweight neurosymbolic framework for hierarchical classification that enforces explicit semantic consistency between parent and child classes at the probability level. By integrating simple inequality-based constraints into standard training pipelines, the proposed approach ensures that model predictions respect known taxonomic structure by construction.

Through controlled ablations and cross-backbone evaluations, we demonstrated that semantic consistency can be substantially improved without degrading predictive performance. Moreover, semantic auditing using representational similarity and error depth metrics shows that consistency-enforced models exhibit more rational failure modes, confusing semantically related classes rather than producing implausible predictions.

Importantly, our approach does not rely on post-hoc explanations or complex symbolic inference engines. Instead, interpretability emerges intrinsically from the logical structure imposed on the model's output space, enabling transparent inspection and auditing of predictions. This makes the proposed framework particularly suitable for safety-critical and decision-support applications where semantic validity and trustworthiness are essential.

Future work will explore extending these constraints to deeper hierarchies, multi-label taxonomies, and domain-specific ontologies, as well as integrating richer forms of symbolic knowledge beyond simple parent-child relationships.

REFERENCES

Bertinetto, L., Mueller, R., Tertikas, K., Samangooei, S., and Lord, N. A. (2020). Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12506–12515. 3

Besold, T. R., Garcez, A. d., Stenning, K., Lamb, L. C., et al. (2017). Neural-symbolic learning and reasoning: A survey and interpretation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 375(2084):20160351. 3

Deng, J., Berg, A. C., and Fei-Fei, L. (2011). Hierarchical semantic indexing for large scale image retrieval. In *CVPR 2011*, pages 785–792. IEEE. 3

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Min-

derer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. 7

Fischer, M., Balunovic, M., Drachsler-Cohen, D., Gehr, T., Zhang, C., and Vechev, M. (2019). Dl2: Training and querying neural networks with logic. In *International Conference on Machine Learning*, pages 1931–1941. PMLR. 3

Garcez, A. d., Besold, T. R., De Raedt, L., Földiák, P., Hitzler, P., Icard, T., Kühnberger, K.-U., Lamb, L. C., Miikkulainen, R., and Silver, D. (2019). Neurosymbolic ai: The 3rd wave. *arXiv preprint arXiv:1905.06088*. 3

Giunchiglia, E. and Lukasiewicz, T. (2020). Coherent hierarchical multi-label classification networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9662–9673. 1, 4

Goren, S., Galil, I., and El-Yaniv, R. (2024). Hierarchical selective classification. *Advances in Neural Information Processing Systems*, 37:111047–111073. 3

Gouidis, F., Vassiliades, A., Patkos, T., Argyros, A. A., Bassiliades, N., and Plexousakis, D. (2020). A review on intelligent object perception methods combining knowledge-based reasoning and machine learning. In *Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice, AAAI-MAKE 2020*, volume 2600. CEUR-WS.org. 3

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. 7

Jin, D., Sergeeva, E., Weng, W.-H., Chauhan, G., and Szolovits, P. (2022). Explainable deep learning in healthcare: A methodological survey from an attribution view. *WIREs Mechanisms of Disease*, 14(3):e1548. 2

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMlR. 5

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. *Technical report, University of Toronto*. 7

Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. (2019). Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32. 3

Kwon, H., Jang, J., Kim, J., Kim, K., and Sohn, K. (2024). Improving visual recognition with hyperbolical visual hierarchy mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17364–17374. 3

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022. 7

Marcus, G. (2020). The next decade in ai: four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*. 1

Park, S., Zhang, Y., Yu, S. X., Beery, S., and Huang, J. (2024). Learning hierarchical semantic classification by grounding on consistent image segmentations. *arXiv preprint arXiv:2406.11608*. 4

Shen, W., Zhou, M., Yang, F., Yu, D., Dong, D., Yang, C., and Tian, J. (2019). Interpretable deep hierarchical semantic convolutional neural network for lung nodule classification. *Expert Systems with Applications*, 128:84–95. 3

Silla Jr, C. N. and Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1):31–72. 3

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR. 7

Ullah, I., Jian, M., Hussain, S., Guo, J., Yu, H., Wang, X., and Yin, Y. (2020). A brief survey of visual saliency detection. *Multimedia Tools and Applications*, 79(45):34605–34645. 2

Valmadre, J. (2022). Hierarchical classification at multiple operating points. *Advances in Neural Information Processing Systems*, 35:18034–18045. 1, 3

Wah, C.-Y., Branson, S., Schwing, A., Neeraj, K., and Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. In *California Institute of Technology*. 7

Weber, S., Zöngür, B., Araslanov, N., and Cremers, D. (2024). Flattening the parent bias: Hierarchical semantic segmentation in the poincaré ball. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28223–28232. 3

Wu, C., Tygert, M., and LeCun, Y. (2019). A hierarchical loss and its problems when classifying non-hierarchically. *Plos one*, 14(12):e0226222. 3

Wu, W., Wang, M., et al. (2025). Deep active learning for image hierarchical classification by introducing dependencies and constraints between classes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. Early Access. 3

Xian, Y., Schiele, B., and Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4582–4591. 7

Xu, J., Zhang, Z., Friedman, T., Liang, Y., and Broeck, G. (2018). A semantic loss function for deep learning with symbolic knowledge. In *International conference on machine learning*, pages 5502–5511. PMLR. 3