# Interpretable 3D Neural Object Volumes for Robust Conceptual Reasoning

Nhi Pham[1] Artur Jesslen[2] Bernt Schiele[1] Adam Kortylewski[3,*] Jonas Fischer[1,*]

[1]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[2]University of Freiburg, Germany
[3]CISPA Helmholtz Center for Information Security, Germany
[*]Equal senior advisorship

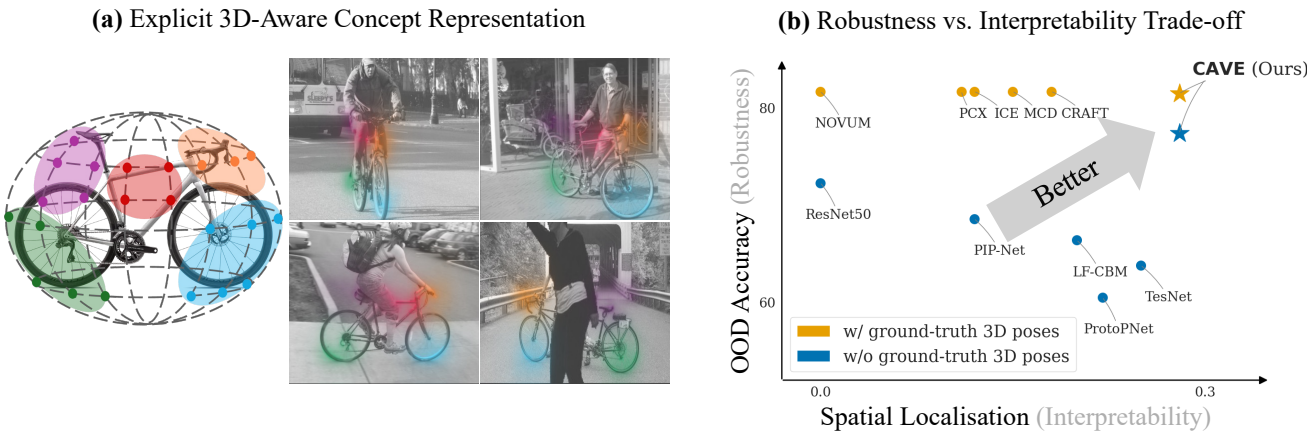


**Fig. 1: CAVE - Concept Aware Volumes for Explanations**. **(a)** We learn 3D object volumes (left), here **ellipsoids**, with concept representations. Each concept captures distinct local features of objects (color coded). At inference (right), these concepts are matched with 2D image features, achieving robust and interpretable image classification. **(b) CAVE achieves the best robustness vs. interpretability tradeoff** across methods (higher is better on both axes). We measure robustness with OOD accuracy (%) on Occluded Pascal3D+ [40], and interpretability with concept spatial localisation (defined in Appendix).

## Abstract

Robustness and interpretability *are crucial to ensure trustworthiness of neural networks. Advances in 3D-aware classifiers that map image features to object volumetric representation, rather than relying on 2D appearance, greatly improve robustness on out-of-distribution (OOD) data. Such classifiers are not yet studied from the perspective of interpretability, while concept-based XAI methods often neglect OOD robustness. We introduce* ***CAVE*** *– **C**oncept **A**ware **V**olumes for **E**xplanations – a new direction that unifies interpretability and robustness in image classification. We design CAVE as a robust, inherently interpretable classifier that learns concepts from 3D object representation. We propose* 3D Consistency (3D-C) *to measure concept consistency. Unlike existing metrics that rely on human-annotated parts, 3D-C uses object meshes as common surfaces to project and compare explanations. CAVE achieves competitive classification accuracy, and discovers consistent concepts across OOD settings.*

## 1. Introduction

Deep neural networks (DNNs) achieve strong performance in various applications, but their predictions often rely on spurious cues and remain opaque [30]. In explainable AI (XAI), post-hoc methods explain pre-trained models without changing their architecture [2, 10, 12, 13, 18], but such explanations are only approximations and need not be faithful to the model's computation. Inherently interpretable models instead impose interpretability during training [1, 6, 27, 28, 31], yet they are often not designed for robustness under distribution shift, e.g., adverse weather (cf. Fig. 2). We introduce **CAVE** (**C**oncept **A**ware **V**olumes for **E**xplanations), an OOD-robust and inherently interpretable image classifier. CAVE builds on neural object volume (NOVs) [22], replacing dense volumetric features with a sparse set of high-level concepts that are more model-faithful (cf. Fig. 1). It leverages zero-shot pose estimates


Figure 2. **CAVE (Ours) discovers consistent concept for Motorbike images under challenging OOD nuisances in OOD-CV dataset.** Columns correspond to inputs with different OOD nuisances: *underwater*, *fog*, *shape*, and *context*. Rows show attributions from NOVUM + ICE (best post-hoc), TesNet (best ad-hoc) and Ours. CAD mesh (right) visualises the **class-level 3D consistency** of a concept, where highlighted regions visualise the aggregated concept attributions across test images. **CAVE** produces more consistent and localised explanations, NOVUM + ICE and TesNet detect concepts inconsistently under nuisances.

from Orient-Anything [43], removing the need for 3D pose supervision, and adapts layer-wise relevance propagation (LRP) for attributing concepts in volumetric architectures. We propose **3D consistency (3D-C)**, a part-annotation-free metric that evaluates whether concepts remain spatially consistent across viewpoints and OOD nuisances by projecting them onto a shared 3D surface. In summary, our main **contributions** include:

(i) CAVE as a robust, inherently-interpretable classifier through ellipsoid NOVs. Our concepts are spatially-aware, and its explanations are model-faithful,

(ii) an adaptation of LRP for concept attribution in classifiers with volumetric representations,

(iii) and a novel part-annotation-free consistency metric 3D-C that captures the spatial coherence of concepts across viewpoints and OOD nuisances.

## 2. Related Work

**Leveraging 3D supervision.** 3D information is useful for 2D feature representations in downstream tasks like segmentation and depth estimation, but these require rich multi-view data [15, 19, 45]. Recently, NOVUM pioneers using 3D information for robust classification, by considering 3D pose information to fit cuboid NOVs to an image [22]. This line of work forms the basis of our approach.

**Concept-based explanations.** A major line of work in XAI focuses on discovering *concept representations*. Post-hoc concept extraction methods such as CRAFT [13, 14] and ICE [46] factorise model activations to uncover latent concepts, while MCD [39] uses sparse subspace clustering to identify concept subspaces, and PCX [9] learns concepts from relevance maps. These approaches offer *implicit interpretability*, and only approximate its computation (i.e., not model-faithful). A different class of approaches makes the model predictions themselves *explicitly interpretable* by design such as concept bottleneck models (CBMs) [23, 31] and prototype-based networks [6, 28, 42].

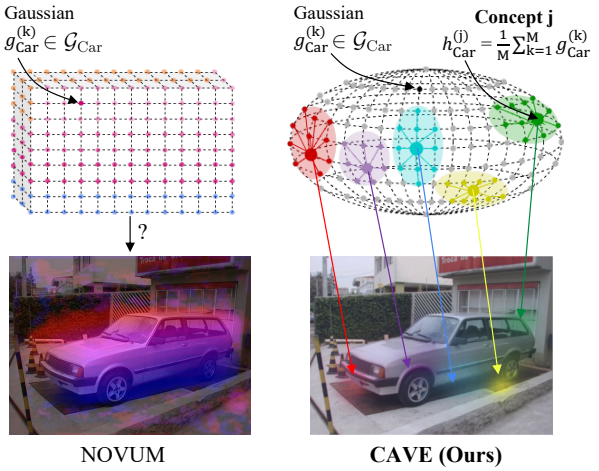


Figure 3. **CAVE** adopts ellipsoid NOVs and produces a sparse set of concepts that replace the dense thousands of Gaussians in NOVUM, thus providing more interpretable explanations.

## 3. Concept-Aware Volumes for Explanations

Our goal is to build an image classifier with two key properties: (i) OOD robust classification, and (ii) inherently interpretable model predictions. While specific solutions exist for each property individually, combining them remains far from trivial. Building upon NOVUM, our method leverages volumetric object representations to simultaneously achieve both robustness and interpretability. We provide a preliminary section on NOVUM in Appendix. We show how to extract a sparse set of interpretable concepts from dense Gaussian features on NOVs, which then form our concept-based NOVs for inherently interpretable classification. Figure 4 gives an overview of CAVE. We attribute these concepts from the model prediction, through these concept-based NOVs, to the input image for explanations using our modified LRP. In Appendix, we discuss how to improve learning NOVs via more expressive shapes and weak 3D supervision with estimated poses for CAVE, thus extending its applicability to settings without ground-truth 3D pose annotations.

**From NOVs to Concept-Based NOVs.** To achieve an inherently interpretable NOV-based classifier, we identify a meaningful concept basis from each NOV and replace the latter with these concepts (cf. Fig. 4). Formally, for a NOV $\mathcal{G}_y \in \mathbb{R}^{K \times C'}$ of class $y$, we formulate our class-wise concept extraction problem with dictionary learning [14, 25]:

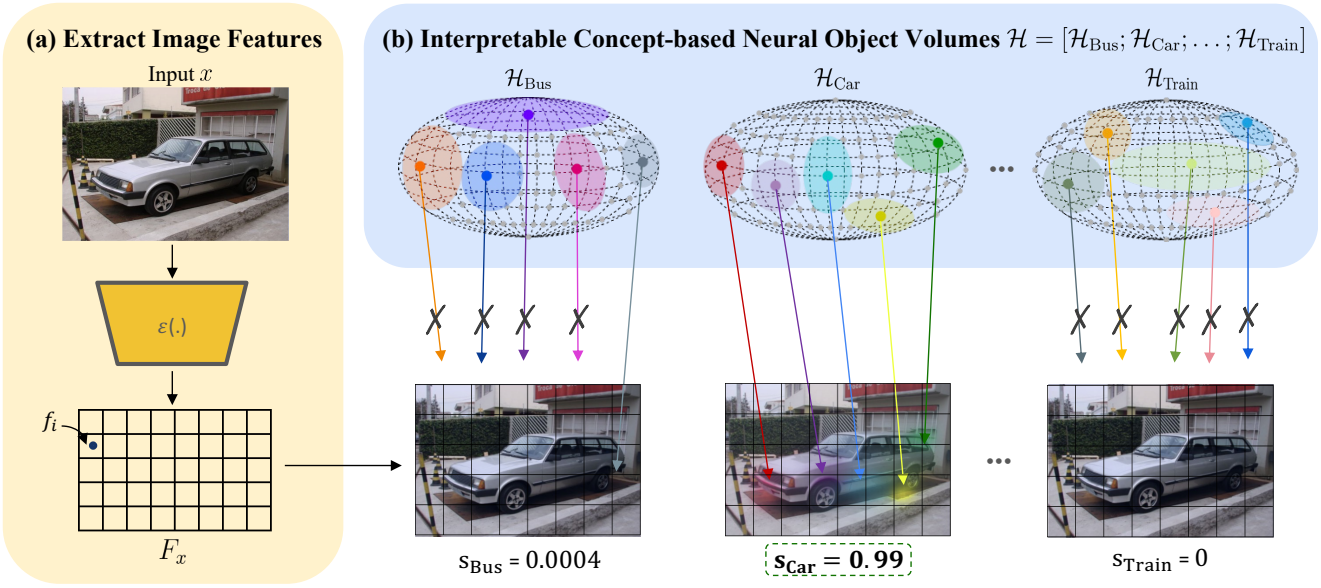


Figure 4. **CAVE – Concept Aware Volumes for Explanations**, a framework for robust conceptual reasoning and classification through 3D-aware **concept-based neural object volumes (NOVs)**. In this visual illustration, colors indicate the top-5 concepts within each class. For classification, CAVE combines **(a)** extracted image features $F_x$ and **(b)** interpretable concept-aware NOVs $\mathcal{H}$ through a bag-of-word concept matching **(c)** with equation 1, where each feature $f_i \in F_x$ is best aligned with $\mathcal{H}$ by cosine similarity. Correct classification happens when image features activates Car concepts, while concepts in other classes fail to align with any feature (crossed-out arrows).

$$(\mathcal{W}_y^\star, \mathcal{H}_y^\star) = \arg\min_{\mathcal{W}_y, \mathcal{H}_y} \|\mathcal{G}_y - \mathcal{W}_y \mathcal{H}_y^\top\|_F^2$$

where the weight matrix $\mathcal{W}_y^* \in \mathbb{R}^{K \times D}$ and the dictionary of $D$ concept vectors $\mathcal{H}_y^* = [h_y^{(1)}, \ldots, h_y^{(D)}]^T \in \mathbb{R}^{D \times C'}$ minimize the element-wise distance between our Gaussian features $\mathcal{G}_y$ and $\mathcal{W}_y \mathcal{H}_y^T$. In the case of hard clustering, the weight matrix $\mathcal{W}_y^*$ reduces to a discrete assignment matrix. This allows clustering to be much more interpretable than methods with less sparse weight matrices. We adopted K-Means clustering for its balance of accuracy, concept sparsity, and alignment to the learned NOVs. The concept dictionary $\mathcal{H}_y^*$ is now seen as a *sparse and interpretable* **concept-based NOV** to replace the original dense NOV $\mathcal{G}_y$ (Fig. 3). We modulate the original feature matching $\phi(F_x, \mathcal{G})$ in NOVUM with **concept matching** $\phi(F_x, \mathcal{H})$ that establishes correspondences between $F_x$ and new volumetric representation $\mathcal{H} = [\mathcal{H}_1^*; \mathcal{H}_2^*; \ldots; \mathcal{H}_N^*] \in \mathbb{R}^{ND \times C'}$. Class $y$ logit is then computed as:

$$s_y = \phi(F_x, \mathcal{H}_y) = \sum_i \max_{j \le D} f_i \cdot h_y^{(j)} \tag{1}$$

This reformulation, illustrated in Fig. 4b-c, enables feature matching against a compact and interpretable concept set instead of thousands of Gaussians, yielding sparser representations, stronger robustness, and more confident predictions compared to NOVUM (cf. Fig. 5).

**Attributing concepts with NOV-aware LRP.** We aim to provide interpretable explanations on the input-level for our NOV-based concepts $\mathcal{H}$, thereby demonstrating the model's reasoning through neural volumetric concepts. To achieve this, we build on LRP, a well-established attribution method that traces relevances from the model's prediction back to the input pixels [3, 32]. A key principle of LRP is the conservation property, which requires the total relevance to remain constant throughout the network [32]. However, we empirically find that when directly applied to NOV-based architectures, LRP fails to uphold this property and instead unfaithfully leaks relevances (see Appendix). We address this by introducing a redistribution rule that preserves the conservation property through the concept-matching operator $\phi(F_x, \mathcal{H})$, ensuring that the total relevance assigned to input pixels equals that at the concept level $\sum_{f_i \in F_x} R_{f_i} = \sum_{h \in \mathcal{H}} R_\phi(h) = R_{y^*}$. This NOV-aware extension allows us to correctly attribute predictions through volumetric concepts with LRP, enabling robust and reliable concept explanations even under challenging OOD conditions. Full derivation is provided in Appendix.

## 4. Experiments

**Datasets and metrics.** We evaluate CAVE on *in-distribution* Pascal3D+ [44] and ImageNet3D [24], and on *OOD* OccludedP3D+ [40] and OOD-CV [47] for accuracy

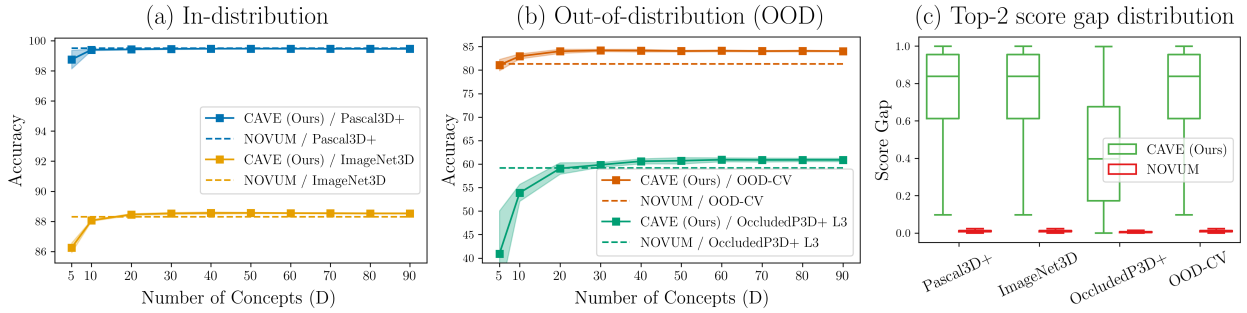


Figure 5. CAVE **replace 1130 dense Gaussians in NOVUM with a compact concept dictionary**, yielding $\sim 98\%$ **sparser representations** that **match or slightly exceed the performance** of NOVUM especially in OOD settings in (a–b). (c) shows improved model prediction; more confident predictions indicate a clearer class separation, which improves reliability [17] and explanation confidence [29].

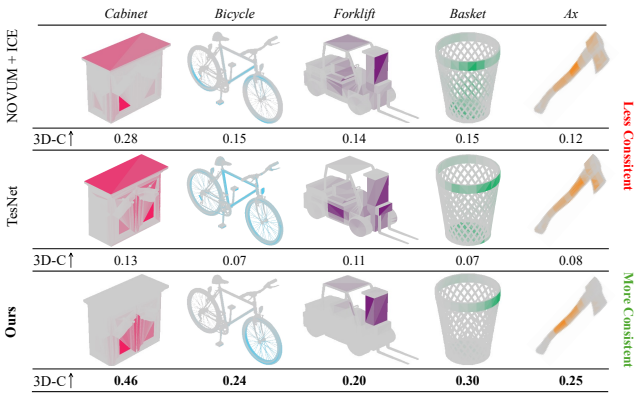


Figure 6. **CAVE (Ours) produces most consistent concepts across different classes**, compared to NOVUM + ICE (best post-hoc) and TesNet (best ad-hoc). Higher 3D-C means more consistent mapping to the same region.

and 3D-C. *Localisation* and *object coverage* are evaluated on Pascal-Part [7]. See further details in Appendix.

**Baselines.** We compare against post-hoc concep methods CRAFT [13], MCD [39], ICE [46], and PCX [9] applied on NOVUM, as well as inherently interpretable baselines LF-CBM [31], ProtoPNet [6], TesNet [42], PIP-Net [28], and MGProto [41]. All methods use ResNet-50 backbone and $D = 20$ concepts per class.

**CAVE discovers spatially consistent concepts**. Table 1 and Fig. 6 show that CAVE achieves the strongest overall concept quality. It yields the best localisation and substantially higher object coverage than all baselines, covering about 80% of the object on average versus 56% for the next best method. CAVE also achieves the highest 3D-C scores across in-distribution and OOD settings, indicating that its concepts remain more stable under viewpoint and nuisance shifts. Even with weak 3D supervision, CAVE outperforms post-hoc explanations built on fully supervised NOVUM.

**CAVE maintains competitive classification accuracies**. We measure robustness in terms of OOD accuracy on OccludedP3D+ and OOD-CV, and further report accuracy on in-distribution Pascal3D+ and ImageNet3D (cf.

| | Models | *Localise.* ↑ | *Coverage* ↑ | *3D Consistency* (3D-C.) ↑ | | | |
|---|---|---|---|---|---|---|---|
| | | Pascal-Part | | Pascal3D+ | ImageNet3D | OccludedP3D+ | OOD-CV |
| Post-hoc | NOVUM + CRAFT [13] | 0.18 | 0.42 | 0.28 | 0.26 | 0.15 | 0.15 |
| | NOVUM + MCD [39] | 0.15 | 0.34 | 0.16 | 0.25 | 0.11 | 0.14 |
| | NOVUM + ICE [46] | 0.12 | 0.44 | 0.28 | 0.27 | 0.15 | 0.15 |
| | NOVUM + PCX [9] | 0.11 | 0.33 | 0.10 | 0.21 | 0.08 | 0.11 |
| Ad-hoc | LF-CBM [31] | 0.20 | 0.56 | 0.15 | 0.14 | 0.13 | 0.11 |
| | ProtoPNet [6] | 0.22 | 0.43 | 0.19 | 0.13 | 0.21 | 0.09 |
| | TesNet [42] | 0.25 | 0.44 | 0.20 | 0.18 | 0.18 | 0.12 |
| | PIP-Net [28] | 0.12 | 0.13 | 0.09 | 0.09 | 0.07 | 0.00 |
| | MGProto [41] | 0.25 | 0.35 | 0.19 | 0.16 | 0.16 | 0.07 |
| | CAVE (Ours) | **0.28** (± 0.001) | **0.80** (± 0.002) | **0.40** (± 0.001) | **0.40** (± 0.001) | **0.23** (± 0.006) | **0.24** (± 0.002) |
| | CAVE (with full 3D supervision) | 0.28 (± 0.001) | 0.87 (± 0.002) | 0.42 (± 0.001) | 0.43 (± 0.0003) | 0.23 (± 0.010) | 0.26 (± 0.001) |

Table 1. **Concept interpretability evaluation**. Our CAVE produces concepts that are spatially localised, sufficiently diverse to cover the object, and robustly consistent across in-distribution and OOD settings. We report our results across 10 random seeds.

| Models | W/o Ground-truth 3D Pose | In-distribution | | Out-of-distribution (OOD) | |
|---|---|---|---|---|---|
| | | Pascal3D+ | ImageNet3D | Occluded P3D+ | OOD-CV |
| LF-CBM [31] | **Yes** | 98.4 | 83.3 | 66.4 | 73.5 |
| ProtoPNet [6] | **Yes** | 97.4 | 74.0 | 60.5 | 71.2 |
| TesNet [42] | **Yes** | 97.6 | 77.9 | 63.8 | 70.1 |
| PIP-Net [28] | **Yes** | 95.7 | 51.0 | 68.6 | 60.0 |
| MGProto [41] | **Yes** | 97.2 | 64.2 | 73.8 | 72.3 |
| CAVE (Ours) | **Yes** | **99.0** (± 0.03) | **84.6** (± 0.02) | **76.8** (± 0.51) | **80.3** (± 0.27) |
| CAVE (with full 3D supervision) | No | 99.4 (± 0.02) | 88.5 (± 0.03) | 81.3 (± 0.30) | 84.0 (± 0.21) |
| NOVUM (with full 3D supervision) | No | 99.5 | 88.3 | 81.7 | 81.3 |

Table 2. **Classification accuracy** ($\%, \uparrow$) **comparison**. CAVE with weak supervision delivers competitive accuracy without ground-truth 3D pose, with only a modest gap to full supervision.

Tab. 2). Across datasets, CAVE with ground-truth 3D poses achieves performance competitive with NOVUM, even slightly surpassing it on large-scale ImageNet3D (+0.2%) and OOD-CV (+2.7%), while using much sparser representations. With weak supervision (no ground-truth 3D poses), CAVE shows comparatively mild drops in performance on ImageNet3D and OOD-CV relative to ground truth pose supervision, yet still clearly outperforms baselines. CAVE provides a unique combination of inherent interpretability and robustness to OOD data unmatched by existing work

## 5. Conclusion

We proposed **CAVE**, a 3D-aware image classifier with concept-based NOVs to jointly achieve OOD robustness and interpretability, while removing the need for ground-truth 3D poses. This enables faithful concepts with strong task performance across OOD settings. We complement existing XAI metrics with our novel 3D-C to measure concept consistency, without relying on pre-defined parts.

## References


[1] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018. 1

[2] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018. 1

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10 (7):e0130140, 2015. 3, 2

[4] Hamed Behzadi-Khormouji and José Oramas. A protocol for evaluating model interpretation methods from visual explanations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1421–1429, 2023. 5, 6

[5] Moritz Böhle, Mario Fritz, and Bernt Schiele. B-cos networks: Alignment is all we need for interpretability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[6] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 1, 2, 4, 6, 13

[7] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014. 4, 5, 11

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[9] Maximilian Dreyer, Reduan Achtibat, Wojciech Samek, and Sebastian Lapuschkin. Understanding the (extra-) ordinary: Validating deep model decisions with prototypical concept-based explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3501, 2024. 2, 4, 6, 13

[10] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631, 2021. 1

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 4

[12] Thomas Fel, Mélanie Ducoffe, David Vigouroux, Rémi Cadène, Mikael Capelle, Claire Nicodème, and Thomas Serre. Don't lie to me! robust and efficient explainability with verified perturbation analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16153–16163, 2023. 1

[13] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2711–2721, 2023. 1, 2, 4, 6, 13

[14] Thomas Fel, Victor Boutin, Louis Béthune, Rémi Cadène, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 9

[15] Stephanie Fu, Mark Hamilton, Laura E. Brandt, Axel Feldmann, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[16] Sheng He, Yanfang Feng, P Ellen Grant, and Yangming Ou. Segmentation ability map: Interpret deep features for medical image segmentation. *Medical image analysis*, 84:102726, 2023. 5

[17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 4

[18] Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. Fast axiomatic attribution for neural networks. *Advances in Neural Information Processing Systems*, 34:19513–19524, 2021. 1

[19] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5693–5702, 2021. 2

[20] Qihan Huang, Mengqi Xue, Wenqi Huang, Haofei Zhang, Jie Song, Yongcheng Jing, and Mingli Song. Evaluation and improvement of interpretability for self-explainable part-prototype networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2011–2020, 2023. 5, 6

[21] Qihan Huang, Jie Song, Jingwen Hu, Haofei Zhang, Yong Wang, and Mingli Song. On the concept trustworthiness in concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21161–21168, 2024. 6

[22] Artur Jesslen, Guofeng Zhang, Angtian Wang, Wufei Ma, Alan Yuille, and Adam Kortylewski. Novum: Neural object volumes for robust object classification. In *European Conference on Computer Vision*, pages 264–281. Springer, 2024. 1, 2, 4, 5, 6, 8, 9, 11, 13

[23] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 2, 13

[24] Wufei Ma, Guofeng Zhang, Qihao Liu, Guanning Zeng, Adam Kortylewski, Yaoyao Liu, and Alan Yuille. Ima-

genet3d: Towards general-purpose object-level 3d understanding. *Advances in Neural Information Processing Systems*, 37:96127–96149, 2024. 3, 4, 5, 6, 10, 11, 12

[25] Julien Mairal, Francis Bach, Jean Ponce, et al. Sparse modeling for image and vision processing. *Foundations and Trends® in Computer Graphics and Vision*, 8(2-3):85–283, 2014. 2

[26] Christoph Molnar. *Interpretable Machine Learning*. 3 edition, 2025. 13

[27] Meike Nauta, Ron Van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14933–14943, 2021. 1

[28] Meike Nauta, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2744–2753, 2023. 1, 2, 4, 6, 13

[29] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s): 1–42, 2023. 4

[30] Fahimeh Hosseini Noohdani, Parsa Hosseini, Aryan Yazdan Parast, Hamidreza Yaghoubi Araghi, and Mahdieh Soleymani Baghshah. Decompose-and-compose: A compositional approach to mitigating spurious correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27662–27671, 2024. 1

[31] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 4, 6, 13

[32] Seitaro Otsuki, Tsumugi Iida, Félix Doublet, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi, and Komei Sugiura. Layer-wise relevance propagation with conservation property for resnet. In *European Conference on Computer Vision*, pages 349–364. Springer, 2024. 3, 2

[33] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019. 5

[34] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. *arXiv preprint arXiv:2001.00396*, 2020. 5

[35] 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. 3

[36] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 3

[37] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In *International Conference on Learning Representations, Workshop Track Proceedings*, 2015. 2, 3

[38] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017. 3

[39] Johanna Vielhaben, Stefan Blücher, and Nils Strodthoff. Multi-dimensional concept discovery (mcd): A unifying framework with completeness guarantees. *Trans. Mach. Learn. Res.*, 2023, 2023. 2, 4, 13

[40] Angtian Wang, Yihong Sun, Adam Kortylewski, and Alan L Yuille. Robust object detection under occlusion with context-aware compositionalnets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12645–12654, 2020. 1, 3, 4, 5

[41] Chong Wang, Yuanhong Chen, Fengbei Liu, Yuyuan Liu, Davis James McCarthy, Helen Frazer, and Gustavo Carneiro. Mixture of gaussian-distributed prototypes with generative modelling for interpretable and trustworthy image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 4, 13

[42] Jiaqi Wang, Huafeng Liu, Xinyue Wang, and Liping Jing. Interpretable image recognition by constructing transparent embedding space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 895–904, 2021. 2, 4, 6, 13

[43] Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. In *Forty-second International Conference on Machine Learning*, 2025. 2, 1, 9, 10, 11, 12

[44] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014. 3, 4, 5, 6, 10, 12

[45] Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. In *European Conference on Computer Vision*, pages 57–74. Springer, 2024. 2

[46] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11682–11690, 2021. 2, 4, 6, 13

[47] Bingchen Zhao, Shaozuo Yu, Wufei Ma, Mingxin Yu, Shenxiao Mei, Angtian Wang, Ju He, Alan Yuille, and Adam Kortylewski. Ood-cv: A benchmark for robustness to out-of-distribution shifts of individual nuisances in natural images. In *European conference on computer vision*, pages 163–180. Springer, 2022. 3, 4, 5, 12

[48] Zhijie Zhu, Lei Fan, Maurice Pagnucco, and Yang Song. Interpretable image classification via non-parametric part prototype learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9762–9771, 2025. 5