Concept-Aware Pruning via Disentangled Subspaces for Robust Convolutional Networks

Kirin Danek
Department of Computer Science
Princeton University
Princeton, NJ
kd9132@princeton.edu

Vikram V. Ramaswamy
Department of Computer Science
Princeton University
Princeton, NJ
vr23@princeton.edu

Abstract

Deep Neural Networks (DNNs) are often pruned to reduce costs in storage, compute, and energy. Existing pruning methods are prone to removing complicated but valid decision strategies in favor of easy but spurious shortcuts. Our novel framework, Concept-aware Network Pruning (CNP), mitigates this shortcut learning by identifying and removing human-interpretable concepts within a model during pruning. CNP augments a pretrained CNN with a virtual concept layer in which nodes represent semantic concepts, ablates human-identified spurious concepts, and then applies attribution-based pruning so that filters serving the ablated concepts are naturally irrelevant. We demonstrate on ImageNet binary classification tasks that CNP can remove spurious watermark reliance while maintaining or improving overall accuracy.

1. Introduction

Deep neural networks (DNNs) are extremely popular due to their impressive performance, leading to their almost ubiquitous use across a variety of devices, including mobile devices where computational resources are scarce. This has led to model "pruning", where several parameters can be removed without significant decreases in accuracy [5, 9, 10, 13–19, 21, 23–25]. Practically, pruned models also allow for lower consumption of resources and are more accessible to people across the globe.

However, existing pruning methods often do not consider the *semantic meaning* of the information that they remove from a neural network. Because of this, pruned models may lose important knowledge or over-rely on potentially spurious decision strategies to make their predictions, leading to less reliable and more biased predictions [5, 15, 23]. Similarly, multiple copies of the same decision strategy (e.g., "green color") may exist in a model, and may simultaneously be kept at the expense of semantically different decision strategies [24].

One way to identify semantic meaning of different parts of a network is through concept-based explanation (C-XAI) frameworks [2, 6–8, 11, 12, 26], which isolate different semantic concepts within the network, either through neurons, vectors or subspaces. We use *disentangled relevant subspace analysis* (DRSA) [3], which outputs orthogonal subspaces corresponding to different concepts. Building on the DRSA framework, we introduce Concept-aware Network Pruning (CNP), a novel pruning framework that avoids this Out-Of-Distribution (OOD) degradation by considering the human-understandable meaning of a model's decision strategies during pruning. As a proof-of-concept, we provide an implementation using disentangled concept subspaces as a pruning intervention and demonstrate on binary classification tasks within ImageNet [4] that CNP can identify and remove spurious watermark concepts, improving OOD robustness while maintaining overall accuracy.[1]

2. Related Work

Explainability-based neural network pruning has been explored in several prior works. Yeom et al. (2019) first used Layer-wise Relevance Propagation (LRP) to assign importance scores to individual network units (filters or weights) as a pruning criterion [25], and Hatefi et al. (2024) later optimized this approach and extended it to Transformers [10]. However, these methods prune units independently without considering semantic meaning of what is being removed, which Yao et al. (2021) identify as a key weakness: redundant filters may be kept while unique ones are discarded, leading to steep accuracy drops at high pruning rates [24]. Yao et al.'s Interpretability-based Filter Pruning (IFP) partially addresses this by using XAI to identify and remove semantically redundant filters within a single layer [24]. Lin et al. (2022) [15] and Wu et al. (2022) [23]

[1] Our code is available at github.com/KirinDanek/NCP.v2

note that pruning disproportionately degrades performance on underprivileged groups, and use per-group importance scores to reduce demographic disparities amplified during compression.

Separately, Concept-based XAI (C-XAI) has developed a range of methods for associating human-interpretable concepts with internal network representations, from neuron-level approaches [2, 7, 11] to broader activation-space methods [6, 8, 12, 26]. Most relevant to our work is the disentangled subspace framework of Chormai et al. (2024), which projects layer activations onto orthogonal concept subspaces via a virtual layer, constructing a one-to-one mapping from semantic concepts to latent subspaces that enables concept attribution and removal [3]. A similar virtual layer mechanism has been used for time series models [22], suggesting its generality as a tool for concept-level intervention. Our framework, Concept-aware Network Pruning (CNP) is the first to bridge C-XAI and attribution-based pruning by using virtual layer representations to identify and suppress concepts based on semantic meaning as an additional consideration to existing pruning strategies.

3. Methodology

We propose Concept-aware Network Pruning (CNP), a three-stage framework that leverages disentangled concept representations to guide neural network compression. By first removing the influence of semantically unimportant (or spurious) concepts from the network's decision strategy, downstream attribution-based pruning naturally identifies and removes the filters that contributed to those concepts, yielding targeted compression.

Stage 1: Augment. We augment a pretrained CNN classifier f with a virtual concept layer inserted after a chosen intermediate layer l. Following Chormai et al. [3], we learn an orthogonal projection matrix $U = (U_1|\cdots|U_K) \in \mathbb{R}^{D\times D}$, where each block $U_k \in \mathbb{R}^{D\times d_k}$ projects the activation vector $\mathbf{a} \in \mathbb{R}^D$ at layer l into a subspace representing concept k. This produces a virtual layer of node activations $\mathbf{h}_k = U_k^\top \mathbf{a}$, where each node is associated with exactly one concept. A reconstructed activation layer $\mathbf{a}' = \sum_{k=1}^{K} U_k U_k^\top \mathbf{a}$ is inserted immediately after the virtual layer and connected to layer $l+1$ via the original weights, ensuring that $\mathbf{a}' = \mathbf{a}$ and the network output is unchanged. The projection weights are learned via Disentangled Relevant Subspace Analysis (DRSA) [3], which optimizes for subspaces that are simultaneously high in relevance for their prototypical instances and comparable to one another in magnitude, using alternating gradient ascent and orthogonalization steps. Relevance for DRSA optimization is computed via Layer-wise Relevance Propagation (LRP) [1], which decomposes the pre-softmax output logit $f_c(X)$ for a target class c into per-node relevance scores by backpropagating relevance layer-by-layer such that $f_c(X) = \sum_h R(h)$ is conserved at every layer. The disentangled property of the virtual layer then allows concept-level relevance to be computed as $R(k) = \sum_{h\in\mathbf{h}_k} R(h)$, with $f_c(X) = \sum_{k\in K} R(k)$.

Stage 2: Ablate. Given the augmented network, we select a subset of concepts $K' \subset K$ to ablate. Target concepts can be identified through two complementary strategies. The first is *latent analysis*: for a candidate spurious attribute (e.g., watermarks), we compute the summed relevance within each subspace as a score and calculate average precision against ground-truth attribute labels to determine whether a subspace is predictive of the spurious attribute. The second is *human evaluation*: an analyst inspects upsampled LRP heatmaps for each concept subspace and judges whether a subspace encodes semantically irrelevant or spurious information [2]. Once target concepts are selected, we ablate them by zeroing out the learned projection weights connecting layer l to the target subspace(s) in the virtual layer—that is, setting $U_{k'} = \mathbf{0}$ for all $k' \in K'$. This yields the modified reconstructed activation $\mathbf{a}' = \sum_{k\in K\setminus K'} U_k U_k^\top \mathbf{a}$, which removes the contribution of the ablated concepts from all downstream computation[3].

Stage 3: Prune. With the ablated concepts no longer influencing the network's output, we apply attribution-based filter pruning following Yeom et al. [25]. Specifically, we use LRP to compute relevance scores for each convolutional filter in the network. Because filters in layers following the augmented layer no longer receive signal that would have propagated through the ablated subspace(s), their activations with respect to the target concept are suppressed during the forward pass. Correspondingly, during the LRP backward pass, relevance flows only through the non-ablated subspaces, so filters that primarily served the ablated concept(s) receive low relevance scores. These filters are therefore naturally identified as low-importance candidates for pruning. Pruning proceeds iteratively: at each step, we rank filters by their LRP relevance and remove a fixed percentage. The augmented layers are then temporarily bypassed for intermediate fine-tuning, allowing the network to recover before re-enabling them for the next round of relevance computation and pruning. Once the target pruning rate is reached, the augmented layers are permanently removed and a final round of recovery fine-tuning is performed.

[2] This in particular differs from past fairness-aware pruning methods in that the training data does not need to have labels for the spurious attribute(s) [15, 23]

[3] An alternative ablation strategy is to propagate only the dataset-average signal through the target subspace(s), which preserves the mean activation statistics while removing instance-specific concept information, removing detected presence of the targeted features at a finer-grained level; we leave investigation of this variant to future work.

Figure 1. **LRP heatmaps and subspaces computed for the basketball class**. The first column shows the original image and the second shows the overall LRP heatmap for the original classifier, while the following 4 columns show the heatmaps of the disentangled subspaces computed. The final column shows the LRP heatmap after using CNP to prune 80% of the network while removing subspace 4 (corresponding to the basketball). Overall, the heatmaps visually rely less on the basketball, and more on the players, their jerseys and the court.

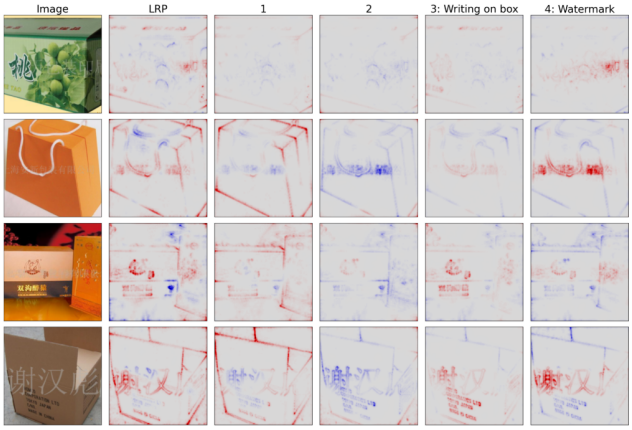

Figure 2. **DRSA concept subspaces extracted from `conv4_3` for the ImageNet "carton" class.** The first column shows the original image, the second shows the overall LRP heatmap, and the following 4 show the heatmaps of the disentangled subspaces computed. The third subspace is "writing on box", while the fourth is "watermark".

4. Evaluation

4.1. Implementation Details

Backbone and task. We use VGG-16 [20] as our backbone architecture. We evaluate on ImageNet [4] in a binary classification setting: for a given target class c with known spurious visual features (e.g., watermarks correlated with the class), we train a binary classifier to distinguish class c from a negative class. Prior to augmentation and ablation, we perform a warmup phase of 15 epochs to train the classification head on the binary task.

Concept extraction. We extract $K = 4$ concept subspaces at layer `conv4_3` of VGG-16 using DRSA [3], learned from 500 positive-class images.

Ablation. Target concepts are identified via the latent analysis and human evaluation strategies described above. Ablation is performed by zeroing out the projection weights $U_{k'}$ for target subspaces $k' \in K'$, effectively removing those subspaces from the virtual layer.

4.2. Results

We evaluate CNP in two settings: a demonstration that confirms concept-targeted pruning removes the intended concept from the network's decision strategy, and a robustness evaluation that tests whether CNP improves out-of-distribution generalization by ablating spurious concepts.

4.2.1. Visual Demo: Basketball

We first demonstrate that CNP successfully removes a targeted concept from a network's internal decision strategy. We apply DRSA to extract $K = 4$ concept subspaces from VGG-16's `conv4_3` layer for the ImageNet "basketball" class [4]. As shown in Figure 1, subspace 4 clearly corresponds to the "ball" concept, as indicated by its LRP heatmaps highlighting the basketball in input images.

We then ablate subspace 4 and prune the network via CNP. The last column in Figure 1 shows standard LRP heatmaps for the CNP-pruned network. The CNP-pruned network exhibits a clear reduction in relevance attributed to the ball region, confirming that the pruning process has effectively diminished the network's reliance on the targeted "ball" concept. Relevance is redistributed to other features such as jerseys, faces, and court markings, indicating that the remaining decision strategy has shifted away from the ablated concept as intended.

4.2.2. Robustness Demo: Imagenet Watermarks

We next evaluate whether CNP can improve OOD robustness by ablating spurious concepts. Many ImageNet classes contain images with visible watermarks from stock photography sources [4]. A model trained on these images may learn to associate watermarks with the positive class, leading to degraded performance on watermark-free positive-class images or, conversely, false positives on negative-class images that happen to contain watermarks. In particular, we consider two classes "carton" and "crate" that have a high rate of watermarks to investigate.

Carton vs. Dugong. We train a binary classifier to distinguish the ImageNet classes "carton" and "dugong." DRSA extracts $K = 4$ subspaces from `conv4_3` for the "carton" class. As shown in Figure 2, subspace 4 is predictive of watermark presence, as its heatmaps highlight watermark regions rather than semantic content of the carton class.

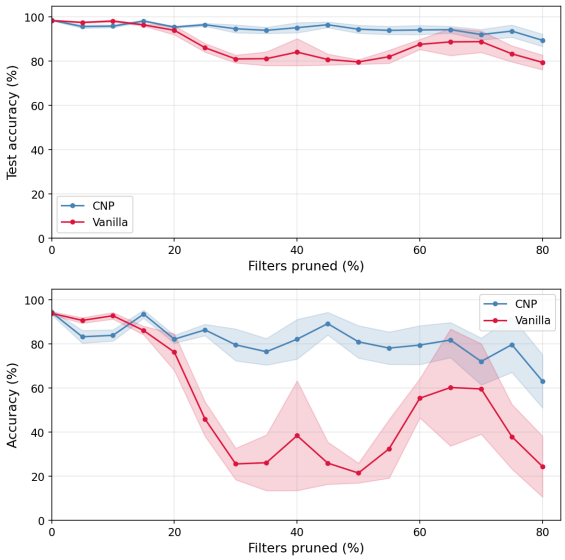

Figure 3. **Performance of pruned networks for carton vs. dugong (mean ± std over $n = 5$ runs).** (*top*) overall accuracy on a balanced test set. (*bottom*) accuracy on the OOD subset (negative-class images with watermarks). CNP maintains higher overall accuracy and improved OOD robustness compared to vanilla LRP pruning.

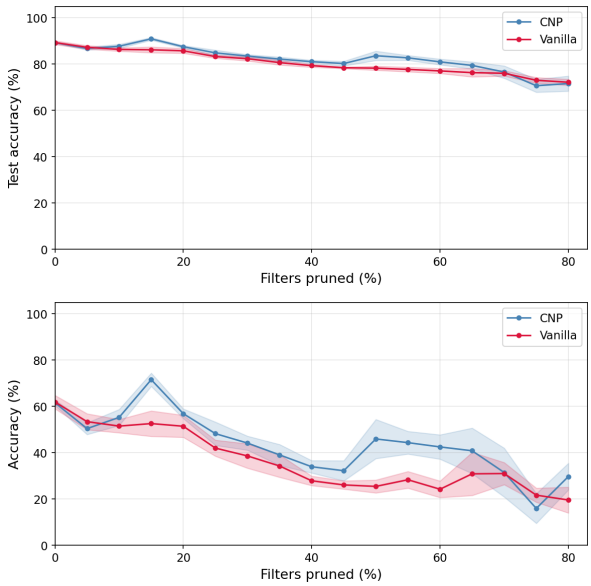

Figure 4. **Performance of pruned networks for crate vs. packet (mean ± std over $n = 5$ runs).** (*top*) overall accuracy on a balanced test set. (*bottom*) OOD subset accuracy (negative-class images with watermarks). The vanilla unpruned model already exhibits low OOD accuracy (∼60%), indicating heavy reliance on the spurious watermark cue. CNP does not significantly improve OOD robustness in this harder setting.

We ablate this watermark subspace via CNP and compare against vanilla LRP pruning (without concept ablation) across pruning iterations. We report two metrics on a test set balancing prevalence of all four combinations of class and watermark presence: (1) overall test accuracy (Figure 3 (*top*)), and (2) accuracy on the OOD subset—negative-class (dugong) images with watermarks, which are the cases most likely to be misclassified by a watermark-reliant model (Figure 3 (*bottom*)).

For the carton–dugong task, CNP maintains increased overall accuracy relative to vanilla LRP pruning throughout the pruning schedule, while showing improved OOD accuracy on the dugong images with watermarks subset. This suggests that ablating the watermark subspace successfully reduces the model's spurious reliance on watermarks without sacrificing overall discriminative performance.

Crate vs. Packet. We repeat the same experimental setup for a harder binary classification task: "crate" vs. "packet," a pair of ImageNet classes that are commonly confused by classifiers due to visual similarity. As with the previous task, we identify and ablate a watermark-associated subspace via CNP (Appendix B).

Results are shown in Figure 4. In contrast to the carton–dugong setting, the crate–packet task reveals a limitation of CNP when the classification problem is difficult. Even the unpruned model achieves only approximately 60% accuracy on the c0w1 subset, suggesting that the model relies substantially on the spurious watermark cue to distinguish between these visually similar classes. As a result, while its performance is still comparable to or outperforms vanilla LRP pruning, CNP does not yield the same OOD improvements observed in the easier task. This highlights an important condition for CNP: when the model strongly relies on the spurious concept—either because the task is too difficult or because the model lacks sufficient non-spurious signal—it may re-encode through other subspaces of the model during fine-tuning and evade ablation.

5. Conclusions

Concept subspaces are more meaningful intervenable units than individual neurons or filters: they carry semantic content that attribution scores alone cannot capture. While existing pruning methods quantify how much units contribute to the model's output, they do not distinguish what that contribution represents, leaving them blind to whether they are removing useful or spurious information. We have shown that pruning at the level of interpretable concepts can successfully target spurious features for removal, improving OOD robustness. We hope this work motivates further exploration of semantically meaningful representations as a basis for network compression.

References

[1] 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):1–46, 2015. 2

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. *CoRR*, abs/1704.05796, 2017. 1, 2

[3] Pattarawat Chormai, Jan Herrmann, Klaus-Robert Müller, and Grégoire Montavon. Disentangled explanations of neural network predictions by finding relevant subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7283–7299, 2024. 1, 2, 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: a large-scale hierarchical image database. pages 248–255, 2009. 1, 3

[5] Mengnan Du, Subhabrata Mukherjee, Yu Cheng, Milad Shokouhi, Xia Hu, and {Ahmed Hassan} Awadallah. Robustness challenges in model distillation and pruning for natural language understanding. In *EACL 2023 - 17th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, pages 1758–1770, United States, 2023. Association for Computational Linguistics (ACL). Publisher Copyright: © 2023 Association for Computational Linguistics.; 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023 ; Conference date: 02-05-2023 Through 06-05-2023. 1

[6] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability, 2023. 1, 2

[7] Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks, 2018. 2

[8] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations, 2019. 1, 2

[9] Masafumi Hagiwara. Removal of hidden units and weights for back propagation networks. In *Proceedings of the International Joint Conference on Neural Networks*, pages 351–353. Publ by IEEE, 1993. Proceedings of 1993 International Joint Conference on Neural Networks. Part 1 (of 3) ; Conference date: 25-10-1993 Through 29-10-1993. 1

[10] Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Reduan Achtibat, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. Pruning by explaining revisited: Optimizing attribution methods to prune cnns and transformers, 2024. 1

[11] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep visual features, 2022. 1, 2

[12] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), 2018. 1, 2

[13] Vadim Lebedev and Victor Lempitsky. Fast convnets using group-wise brain damage. *Arxiv*, 2016. 1

[14] Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*. Morgan-Kaufmann, 1989.

[15] Xiaofeng Lin, Seungbae Kim, and Jungseock Joo. Fairgrape: Fairness-aware gradient pruning method for face attribute classification, 2022. 1, 2

[16] Giosué Cataldo Marinó, Alessandro Petrini, Dario Malchiodi, and Marco Frasca. Deep neural networks compression: A comparative survey and choice recommendations. *Neurocomputing*, 520:152–170, 2023.

[17] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient inference, 2017.

[18] Michael C Mozer and Paul Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. In *Advances in Neural Information Processing Systems*. Morgan-Kaufmann, 1988.

[19] James O'Neill. An overview of neural network compression. *CoRR*, abs/2006.03669, 2020. 1

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv 1409.1556*, 2014. 3

[21] Lucas Theis, Iryna Korshunova, Alykhan Tejani, and Ferenc Huszár. Faster gaze prediction with dense networks and fisher pruning. *CoRR*, abs/1801.05787, 2018. 1

[22] Johanna Vielhaben, Sebastian Lapuschkin, Grégoire Montavon, and Wojciech Samek. Explainable ai for time series via virtual inspection layers. *Pattern Recognition*, 150: 110309, 2024. 2

[23] Yawen Wu, Dewen Zeng, Xiaowei Xu, Yiyu Shi, and Jingtong Hu. Fairprune: Achieving fairness through pruning for dermatological disease diagnosis, 2022. 1, 2

[24] Kaixuan Yao, Feilong Cao, Yee Leung, and Jiye Liang. Deep neural network compression through interpretability-based filter pruning. *Pattern Recognition*, 119:108056, 2021. 1

[25] Seul-Ki Yeom, Philipp Seegerer, Sebastian Lapuschkin, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Pruning by explaining: A novel criterion for deep neural network pruning. *CoRR*, abs/1912.08881, 2019. 1, 2

[26] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2

A. Hyperparameters

Table 1 summarizes the key hyperparameters in the CNP pipeline and the values explored during our search. Hyperparameters were validated on 150 images per subgroup before final selection; all reported results use a held-out class- and subgroup-balanced test set.

Hyperparameter	Values tested	Selected
Num. subspaces K	4, 8, 16	4
Subspace dim. d_k	128, 64, 32	128
DRSA layer l	`conv4_3`	`conv4_3`
LRP variant	ϵ, $\alpha_1\beta_0$, γ	$\alpha_1\beta_0$
Fine-tuning epochs/iter.	0, 2, 5, 10	2
Filter ranking data	pos. only, pos. + neg.	pos. only
Pruning step size	5%, 10%	5%

Table 1. Hyperparameter search space and selected values. K and d_k are inversely coupled such that $K \times d_k$ equals the activation dimensionality at layer l. Zero intermediate fine-tuning epochs caused catastrophic degradation; all nonzero values performed comparably. Additional hyperparameters not searched over include fine-tuning optimization.

A.1. Pruning schedule.

We use LRP-$\alpha_1\beta_0$ with an epsilon stabilizer to preserve the identity mapping of non-ablated subspaces during relevance backpropagation. LRP relevance for filter importance ranking is calculated on a set of 500 positive-class images, with respect to only the positive logit output. Pruning proceeds in increments of 5% of total network filters per step, up to a total pruning rate of 80%. At each pruning step, we perform 2 epochs of intermediate fine-tuning on 650 images from each class. Crucially, we disable pruning of the virtual layer, the reconstructed activation layer, and `conv4_3` throughout the pruning process; freezing `conv4_3` avoids the need to recompute concept subspaces after each pruning iteration.

B. Crate DRSA Subspaces

Figure 5 depicts the subspaces extracted from the crate-packet classification model. We ablate subspace 2 (watermark) during pruning.

C. Additional Limitations

LRP propagation rule hyperparameters (e.g., $\alpha - \beta, \gamma, \epsilon$) are architecture-dependent and require careful tuning, rendering usage on networks with variable-sized layers very difficult. DRSA's subspace disentanglement is imperfect: the orthogonal projection assumes concepts occupy non-overlapping linear subspaces, which is an approximation. In particular, visual features that tend to co-occur in the same spatial locations (e.g., blond hair and light skin) are difficult to separate into distinct subspaces, limiting the precision of targeted ablation. In practice, obtaining perfectly disentangled concept representations is not yet solved.

To ensure concepts do not re-encode through other subspaces of the chosen layer during intermediate fine-tuning, DRSA subspaces should be recomputed after every intermediate fine-tuning session; however, this was too costly for our experiments.

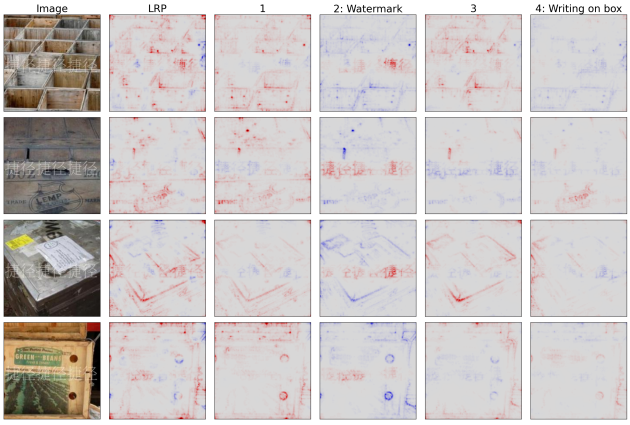

Figure 5. **DRSA concept subspaces extracted from `conv4_3` for the ImageNet "crate" class.** The first column shows the original image, the second shows the overall LRP heatmap, and the following 4 show the heatmaps of the disentangled subspaces computed. The fourth subspace is "writing on box", while the second is "watermark".