Multi-Granularity Concept Whitening for Neural Network Interpretability

Russell Barton, Viet Hung Le, Yunhong Shan, Alexander Katopodis, Jon Donnelly, Eric Chen, Chaofan Chen, Cynthia Rudin
Duke University; University of Maine | CVPR 2026 Workshop on Explainable AI for Computer Vision

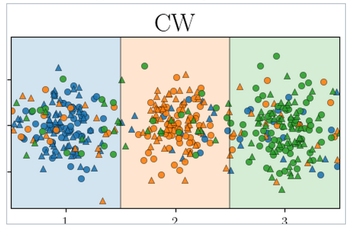
CW makes internal CNN axes interpretable. MGCW makes those axes hierarchical.

CVPR JUNE 3-7, 2026 DENVER COLORADO

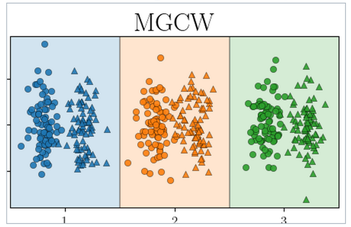
1. Motivation
fine-grained concepts are not flat

Concept Whitening aligns axes with concepts, but it does not know that *wing-blue*, *nape-blue*, and *tail-blue* belong to different semantic regions.

Flat CW: global competition



MGCW: local high-level subspaces



MGCW partitions the latent space into high-level concept subspaces, so fine-grained concepts compete locally rather than globally.

- **Flat CW:** every concept competes in one global rotation.
- **MGCW:** each high-level concept gets an axis slice.
- **Result:** color and shape attributes are interpreted relative to the right part.

Why CUB is the stress test
Bird attributes naturally factor into part and subtype: nape color, wing color, beak shape, tail shape.

2. Method
whitening, rotation, local alignment

Whiten and rotate

$$\psi(\mathbf{z}) = \mathbf{W}(\mathbf{z} - \boldsymbol{\mu}), \quad \mathbf{W}^\top \mathbf{W} = \boldsymbol{\Sigma}^{-1}$$
$$\Psi(\mathbf{z}) = \mathbf{Q}\psi(\mathbf{z}), \qquad \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$$

Whitening *decorrelates features; the orthogonal rotation Q chooses the concept-aligned coordinate system.*

Labeled axes + free axes

$$\max_{\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}} \mathcal{L}_{\text{labeled}} + \lambda \mathcal{L}_{\text{free}}$$

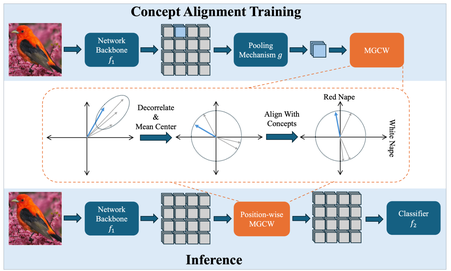
$$\mathcal{L}_{\text{labeled}} = \sum_{k \in \mathcal{Q}_{\text{sub}}} \mathbb{E}_{x \in \mathcal{D}_k^{\text{sub}}} \left[\mathbf{q}_k^\top \psi(\mathbf{z}_x) \right]$$

$$\mathcal{L}_{\text{free}} = \sum_{h} \mathbb{E}_{x \in \mathcal{D}_h^{\text{high}}} \left[\max_{j \in S_h} \mathbf{q}_j^\top \psi(\mathbf{z}_x) \right]$$

Objective. *Labeled examples push known axes. Free examples choose the most active axis inside their parent subspace.*

Labeled
Known subtype, fixed axis: *nape:red* pushes the nape-red axis.

Free
Parent known, subtype unknown: a nape crop updates only a nape-axis winner.

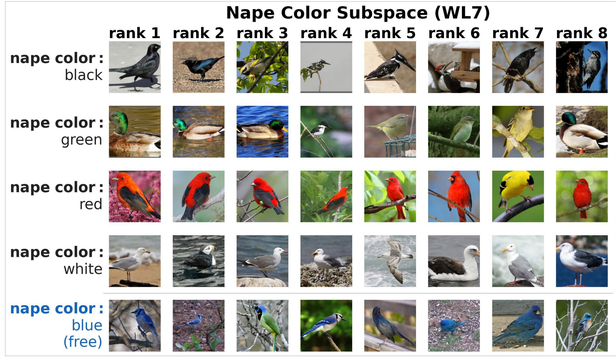


Training/inference. *Classification remains normal; concept alignment updates the rotation using cropped concept images.*

3. Free axes
bounded discovery inside a parent concept

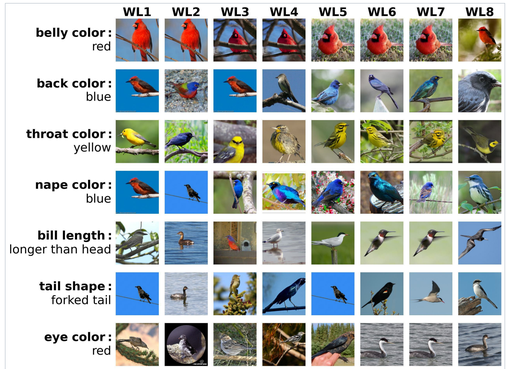
Free axes are reserved slots inside a parent subspace. They let the model use partial labels such as *some nape color* without manually naming every subtype.

Withhold blue nape; ask the free axis to find it



Nape-color subspace, ResNet-18 WL7. *The final row is a free axis. Although blue nape is not given as a labeled sub-concept, the free axis retrieves coherent blue-naped birds.*

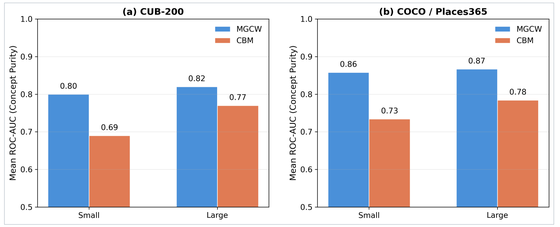
- Discovery is **bounded**: the free axis can specialize only inside the nape subspace.
- This turns weak parent labels into structure instead of letting them contaminate unrelated axes.



Layer placement. *Later MGCW insertion points yield more semantic concept activations.*

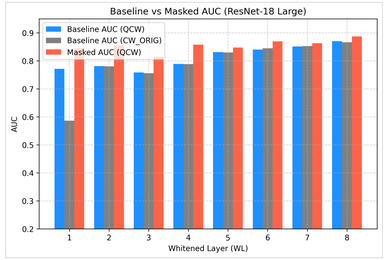
4. Results
purity rises while accuracy stays close

Concept purity



MGCW vs CBM. *Mean ROC-AUC concept purity improves on CUB and Places365 concept sets.*

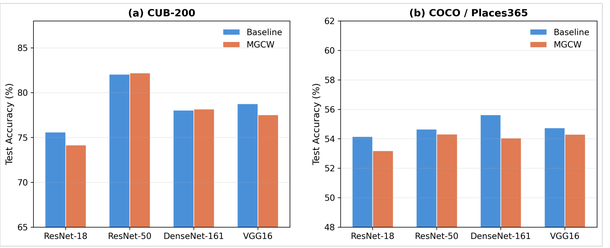
The hierarchy is measurable



Masked AUC. *Restricting evaluation to the correct high-level subspace makes concepts cleaner; CUB Large/RN-18 rises from 0.616 to 0.865.*

Accuracy remains close

Because concept alignment does not train the backbone, accuracy stays near baseline across standard CNN backbones.



Accuracy across backbones. *Selected MGCW insertion layers preserve predictive performance while adding concept structure.*

Key references: Chen, Bei, and Rudin, Concept Whitening, Nature Machine Intelligence, 2020. Koh et al., Concept Bottleneck Models, ICML, 2020. Datasets: CUB-200-2011, COCO, Places365.

Code: github.com/rbarton124/MultiGranularityCW