DCBM: Data-efficient Visual Concept Bottleneck Models

Katharina Prasse* [1] Patrick Knab* [2] Sascha Marton [2] Christian Bartelt [2] Margret Keuper [1,3]

*Equal contribution [1]Data and Web Science Group, University of Mannheim [2]Clausthal University of Technology

[3]Max-Planck-Institute for Informatics, Saarland Informatics Campus

Paper

Code

Motivation

- Concept Bottleneck Models (**CBMs**) learn a linear mapping from concept activations to classes that are inherently interpretable.
- CBMs main objectives:
  → Meaningful **human-interpretable concepts**.
  → Concepts are sufficiently **specific for the given task**.
  → **Efficient extraction of concepts** from training images/classes.

No Description,

No Supervision,

No External Data.

—

Extract Concepts from

YOUR Data.

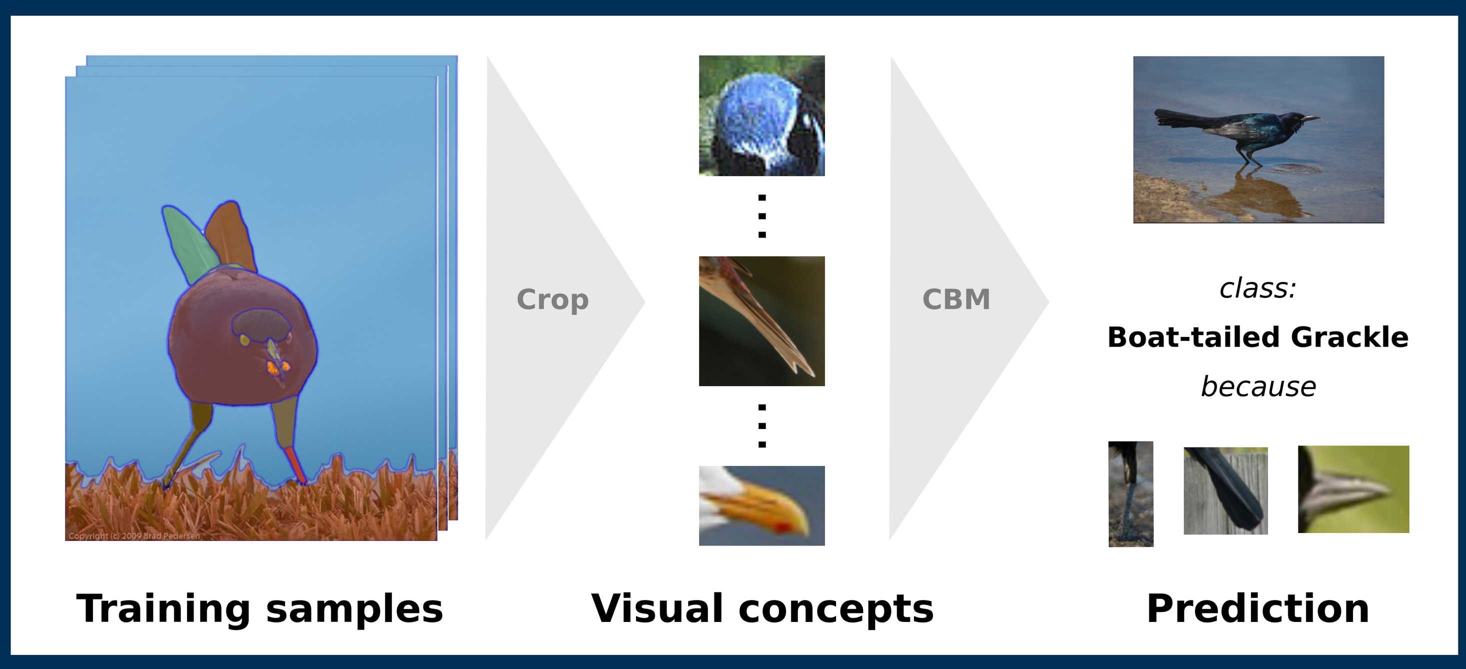

Figure 1: Using vision foundation models, we use cropped image regions as concepts for CBM training. Based on few concept samples (50 imgs/class), DCBMs offer interpretability even for fine-grained classification.

Framework: Data-efficient CBMs

Step 1: Concept proposals are created using **foundation models** for segmentation / detection.

Step 2: Concepts are generated by **clustering concept proposals** to remove redundancies.

Step 3: **CBM is trained** to map concept activations to class labels.

Step 4: **Visual concepts are mapped to text** within CLIP space.

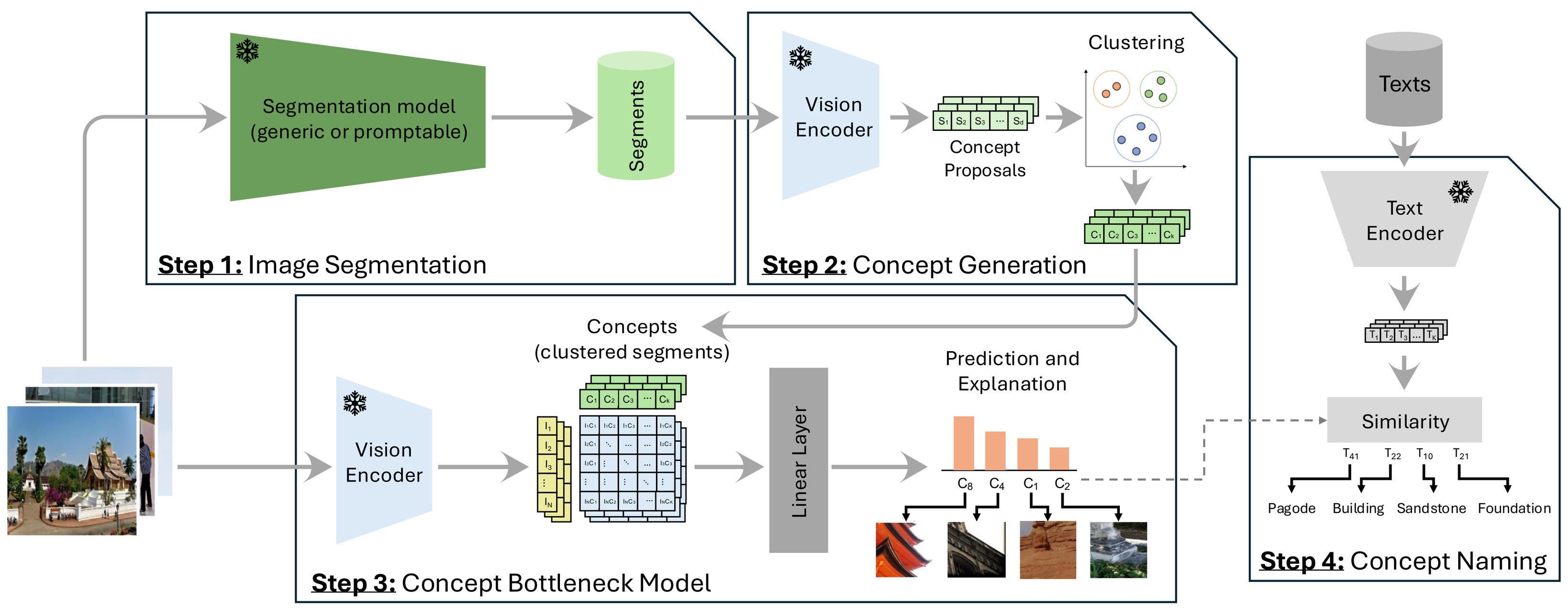

Figure 2. The **DCBM framework** generates concept proposals through foundation models (Step 1). These proposals are clustered (Step 2); the resulting concepts train a sparse CBM (Step 3). Image–text alignment then maps each visual concept to its textual counterpart (Step 4). Undesired concepts can be pruned after Step 2.

Qualitative & Quantitative Results

- DCBMs **perform within at most 6% of the linear probe for all datasets (9).**
- Mask-RCNN concept proposals outperform SAM2 and GDINO.
- DCBM **excels in domain- specific tasks** (e.g., CUB).
- DCBM concepts are applicable in OOD settings.
- DCBMs achieve competitive performance using just 50 imgs/class as concept samples.

Table 1. Top-1 **accuracy comparison** across CBM models.

Model	CLIP ViT L/14				
	IMN	Places	CUB	Cif10	Cif100
Linear Probe ↑	83.9*	55.4	85.7	98.0*	87.5*
Zero-Shot ↑	75.3*	40.0	62.2	96.2*	77.9*
LF-CBM [3] ↑	-	49.4	80.1	97.2	83.9
LaBo [6] ↑	**84.0***	-	-	97.8*	86.0*
CDM [4] ↑	83.4*	55.2*	-	95.9	82.2
DCLIP [2] ↑	75.0*	40.5*	63.5*	-	-
DN-CBM [5] ↑	83.6*	**55.6***	-	**98.1***	**86.0***
DCBM-SAM2 (Ours) ↑	77.9	52.1	**81.8**	97.7	85.4
DCBM-GDINO (Ours) ↑	77.4	52.2	**81.3**	97.5	85.3
DCBM-MASK-RCNN (Ours) ↑	77.8	52.1	**82.4**	97.7	85.6

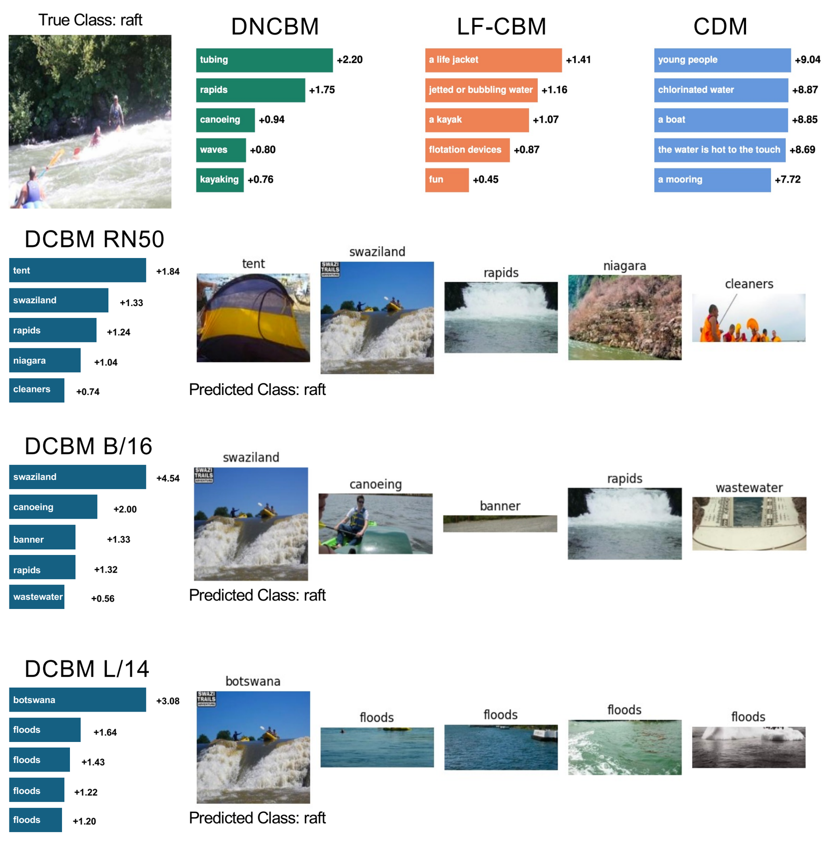

Figure 3. **CBM concept explanation comparison**. DCBM explanations contain no abstract concepts (e.g. fun, chlorinated water).

Table 2. **Data-efficiency.** DCBM concept proposals are generated from 50 imgs per class.

	DN-CBM [5]	DCBM-ImageNet
Dataset	CC3M	50 k images (50 imgs/class)
Mem	850 GB (256×256)	6 GB
No extra data	x	✓

Table 3. **OOD performance.** Error rate changes compared between visual CBMs (CLIP ViT-L/14) on ImageNet-R.

	IN-200	IN-R	Gap(%)
DN-CBM [5] ↓	16.4	55.2	38.8
DCBM-SAM2 (Ours) ↓	21.1	**48.5**	**27.4**
DCBM-GDINO (Ours) ↓	22.6	**47.2**	**24.6**
DCBM-MASK-RCNN (Ours) ↓	22.2	**44.6**	**22.4**

[1] M. Bohle, M. Fritz, and B. Schiele. Convolutional dynamic alignment networks for interpretable classifications. In *CVPR*, 2021.

[2] S. Menon and C. Vondrick. Visual classification via description from large language models. In *ICLR*, 2023.

[3] T. Oikarinen, S. Das, L. Nguyen, and L. Weng. Label-free concept bottleneck models. In *ICLR*, 2023.

[4] K. P. Panousis, D. Ienco, and D. Marcos. Sparse linear concept discovery models. In *ICCV*, 2023.

[5] S. Rao, S. Mahajan, M. Böhle, and B. Schiele. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *ECCV*, 2024. First 2 authors contribute equally.

[6] Y. Yang, A. Panagopoulou, S. Zhou, D. Jin, C. Callison-Burch, and M. Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, 2023.

Funded by the European Union NextGenerationEU