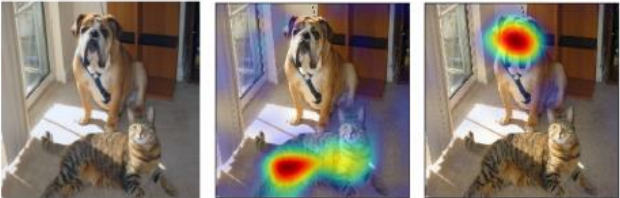
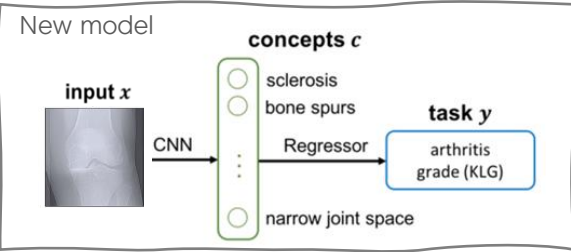
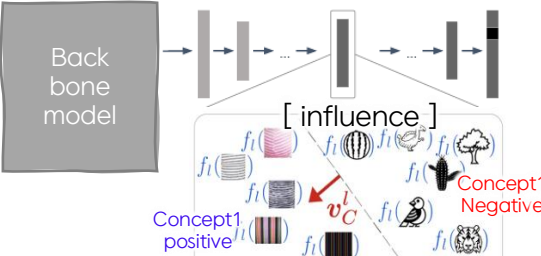
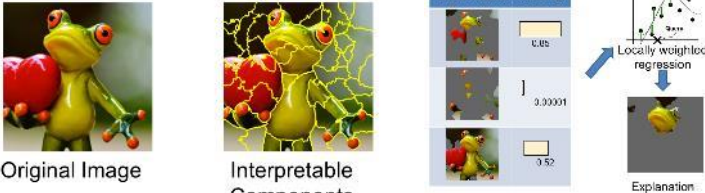
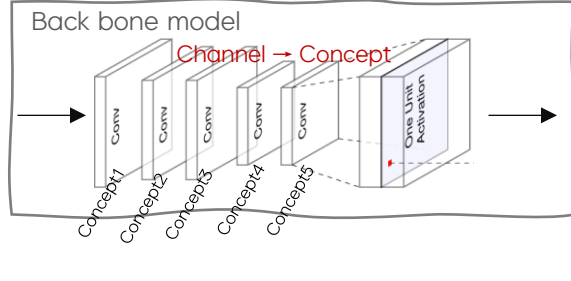
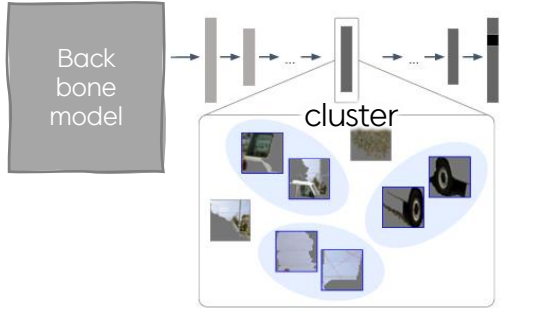
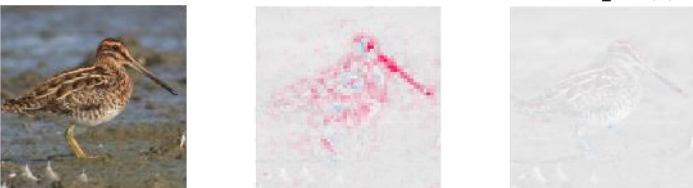


LABEL-FREE **C**ONCEPT **B**OTTLENECK **M**ODELS

ICLR 2023

Feature Attribution vs Concept-based XAI

Feature Attribution XAI	Concept-based XAI
Grad-CAM (2016)  Original Image Grad-CAM 'Cat' Grad-CAM 'Dog'	Supervised CBM (2020)  TCAV (2018) 
LIME (2016)  Original Image Interpretable Components Explanation	Weakly- / Unsupervised Network Dissection (2017)  ACE (2019) 
SHAP (2017~)  Original Image dowitcher red-backed_sandpiper	

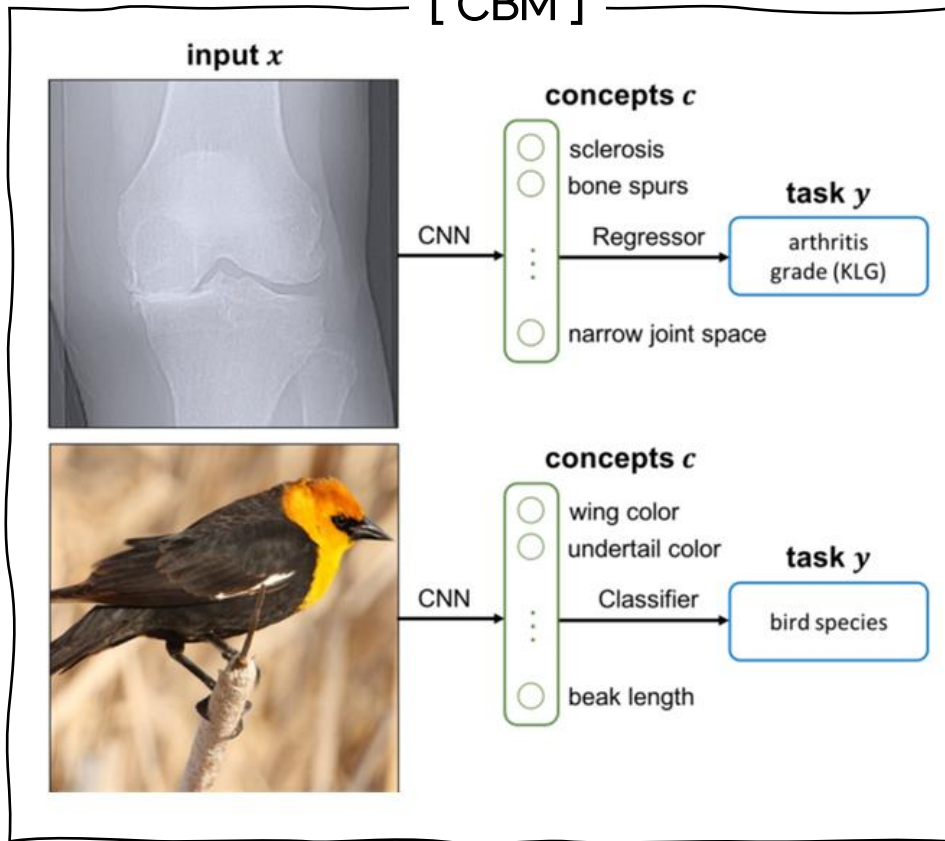
Landscape of Concept-based XAI model in Computer vision

Base	Model	Structure	Additional training	Concept Supervision	Concept Granularity	Concept Form	Model Agnostic	Backbone
Dissection	Network Dissection (2017)	Post-hoc	No	Weakly-supervised	Neuron	Explicit	No	CNN
	GAN-Dissect (2019)	Post-hoc	No	Unsupervised	Neuron	Explicit	No	GAN (generator)
	CLIP-Dissect (2022)	Post-hoc	No	Weakly-supervised	Neuron	Explicit	No	CLIP
CAV	TCAV (2018)	Post-hoc	Yes	Fully-supervised	Image	Semi-Explicit	Yes	CNN
ACE	ACE (2019)	Post-hoc	Yes	Unsupervised	Image	Implicit	No	CNN
	CRAFT (2023)	Post-hoc	Yes	Unsupervised	Image	Implicit	Yes	CNN
CBM	CBM (2020)	Pre-hoc		Fully-supervised	Image	Explicit	No	CNN
	Post-hoc CBM + CAV (2022)	Post-hoc	Yes	Fully-supervised	Image	Explicit	Yes	CNN
	Post-hoc CBM + CLIP (2022)	Post-hoc	No	Weakly-supervised	Image	Explicit	No	CLIP
	Label-Free CBM (2023)	Pre-hoc		Weakly-supervised	Image	Explicit	Yes	CNN

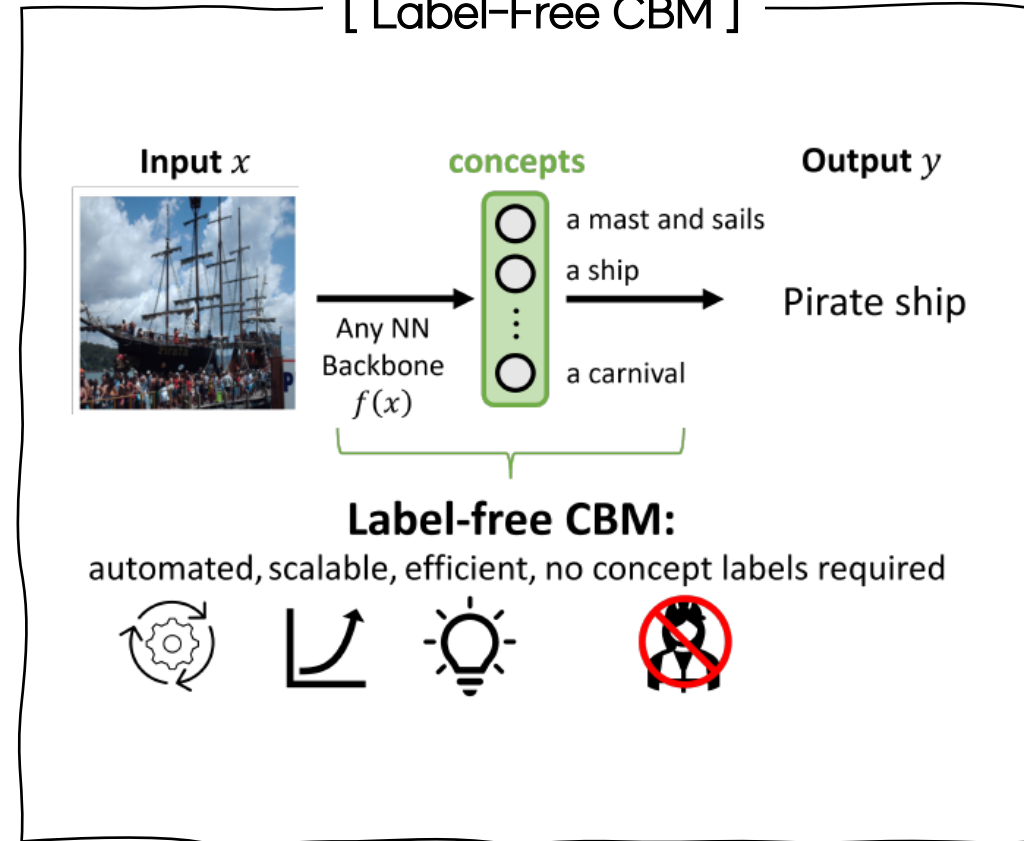
Label-Free CBM

CBM vs Label-Free CBM

[CBM]

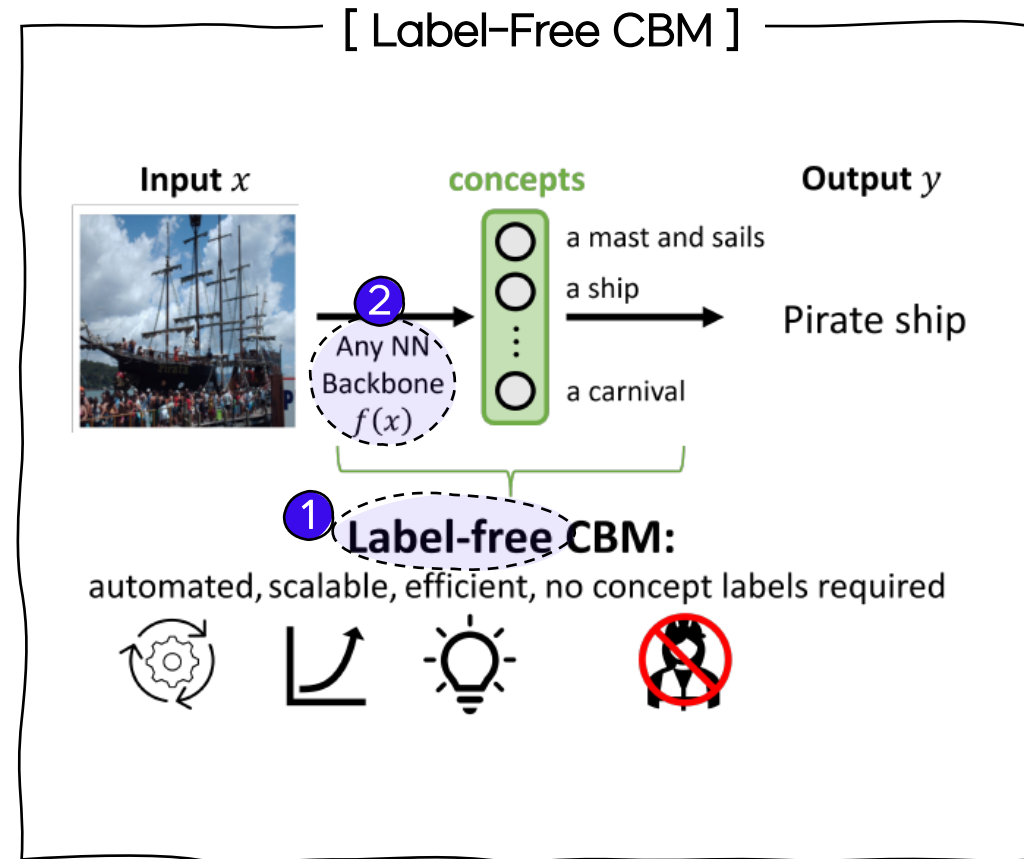
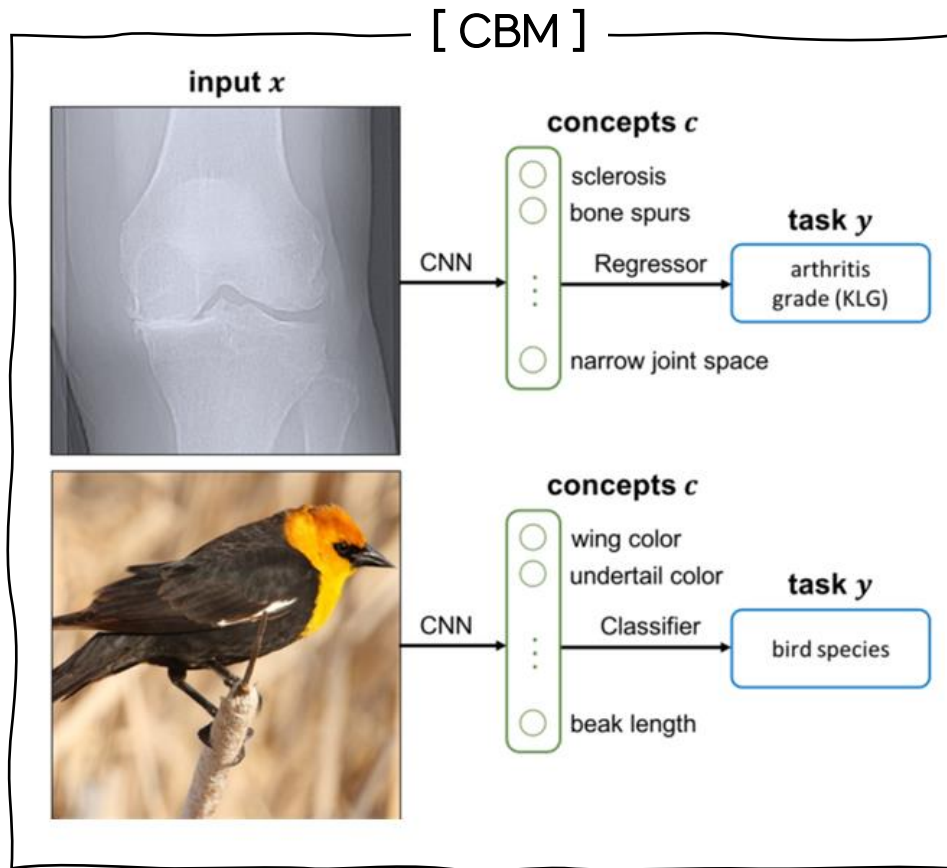


[Label-Free CBM]



Label-Free CBM

CBM vs Label-Free CBM



Architecture

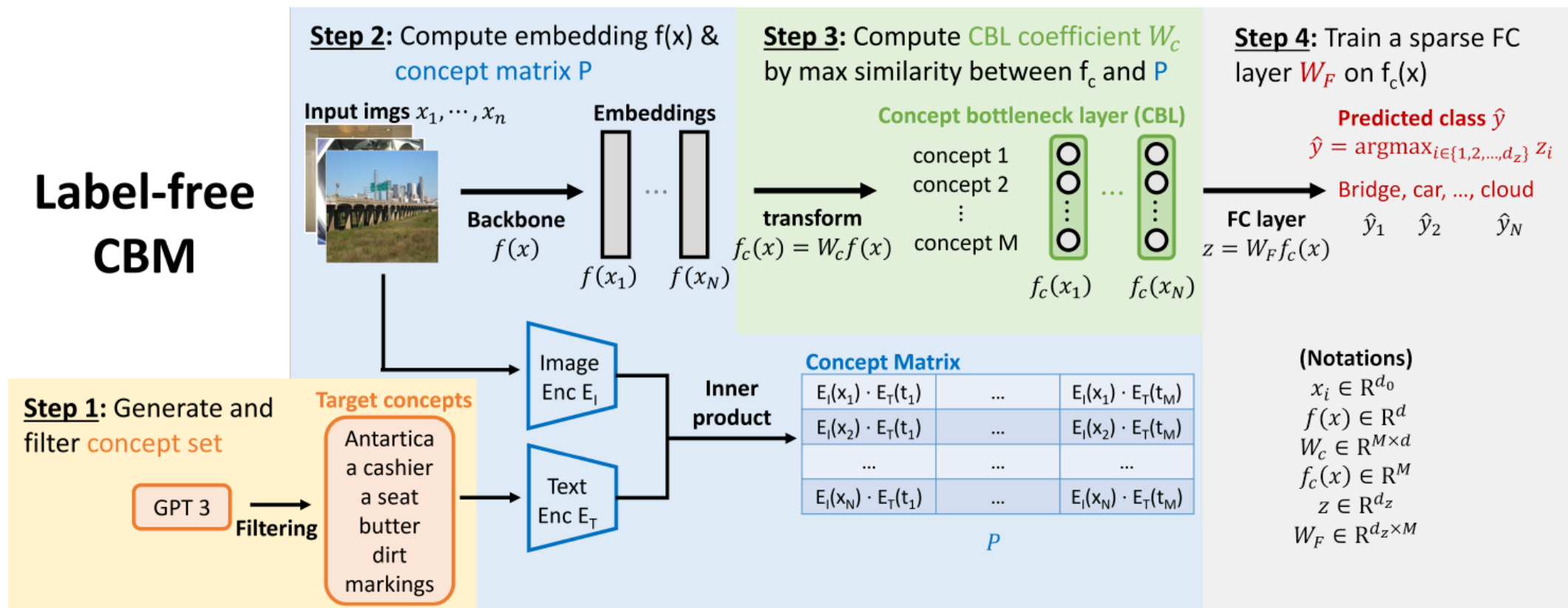


Figure 2: Overview of our pipeline for creating label-free CBM.

Label-Free CBM

Architecture

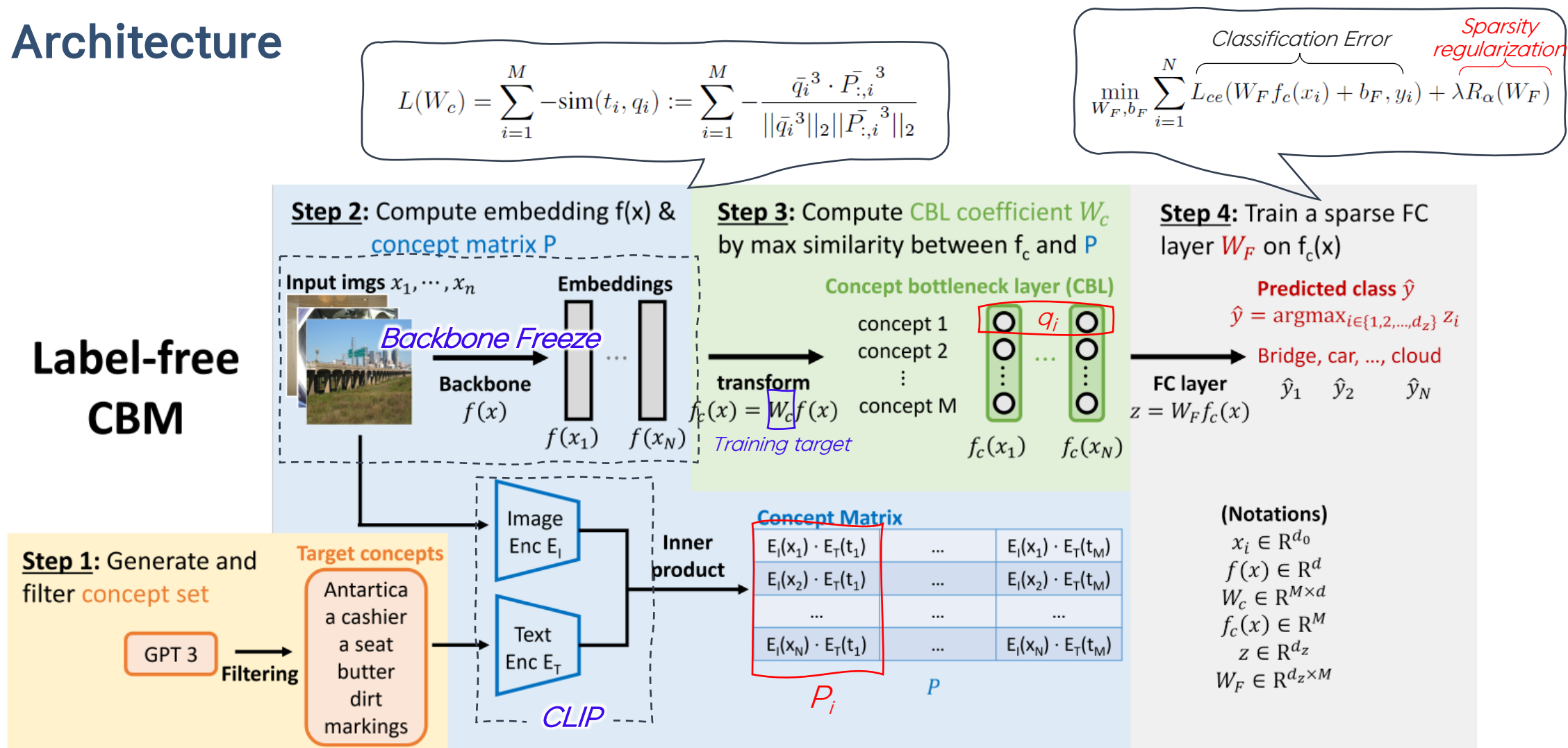


Figure 2: Overview of our pipeline for creating label-free CBM.

Appendix

Experimental setup

Baseline

Model	Concept Usage	Concept Supervision	Structure
CBM	O	Fully-supervised	Pre-hoc
Post-hoc CBM (P-CBM)	O	Fully-supervised	Post-hoc
Post-hoc CBM (w/ Concept Net)	O	Weakly-supervised	Post-hoc
Sparse Linear Model	X	-	Post-hoc
IBD	X	-	Post-hoc
Label-Free CBM	O	Weakly-supervised	Pre-hoc

Datasets

- CIFAR-10
- CIFAR-100
- CUB-200
- Places365
- ImageNet

Evaluation

- (1) Accuracy
- (2) Global Explainability
- (3) Instance-level Explainability
- (4) Manual Editing of Final Layer

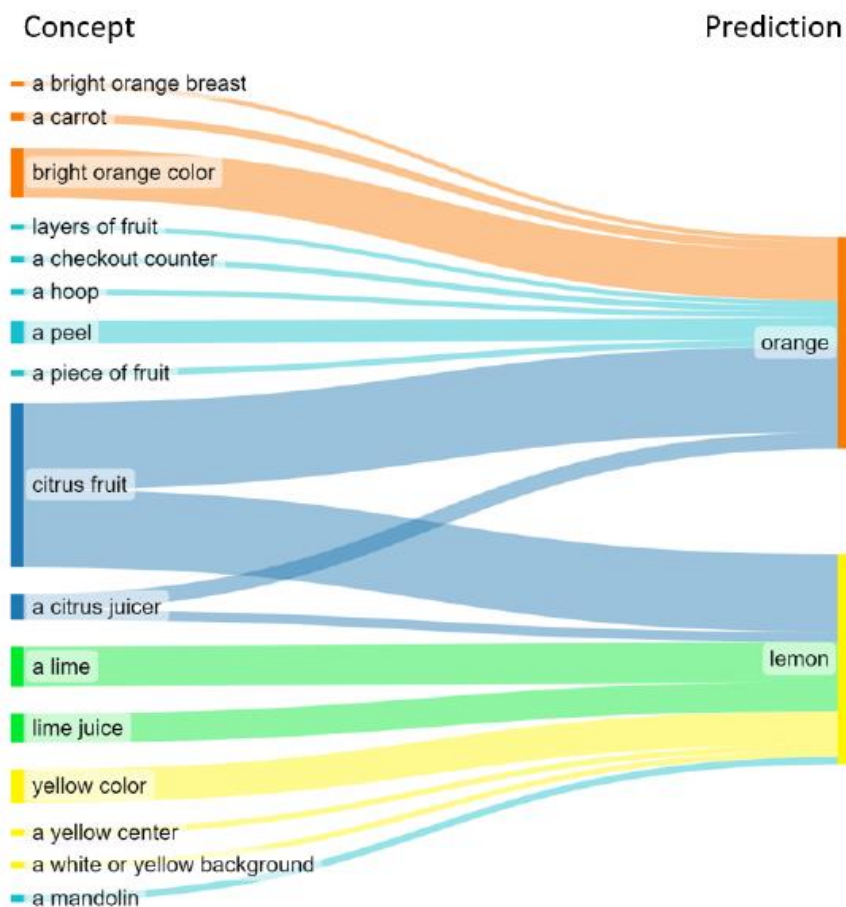
Result

	(I) Flexibility		(II) Interpretability		(III) Performance	
Method:	Without labeled concept data	Any network architecture	Sparse final layer	All features interpretable	Preserves accuracy	Extends to ImageNet scale
CBM	No	Yes	No	Yes	No	No
IBD	No	Yes	No	No	Yes	No
P-CBM	No	Yes	Yes	Yes	No	No
P-CBM (CLIP)	Yes	No	Yes	Yes	No	Maybe
P-CBM-h	No	Yes	Yes	No	Yes	No
P-CBM-h (CLIP)	Yes	No	Yes	No	Yes	Maybe
Label-free CBM (This work)	Yes	Yes	Yes	Yes	Yes	Yes

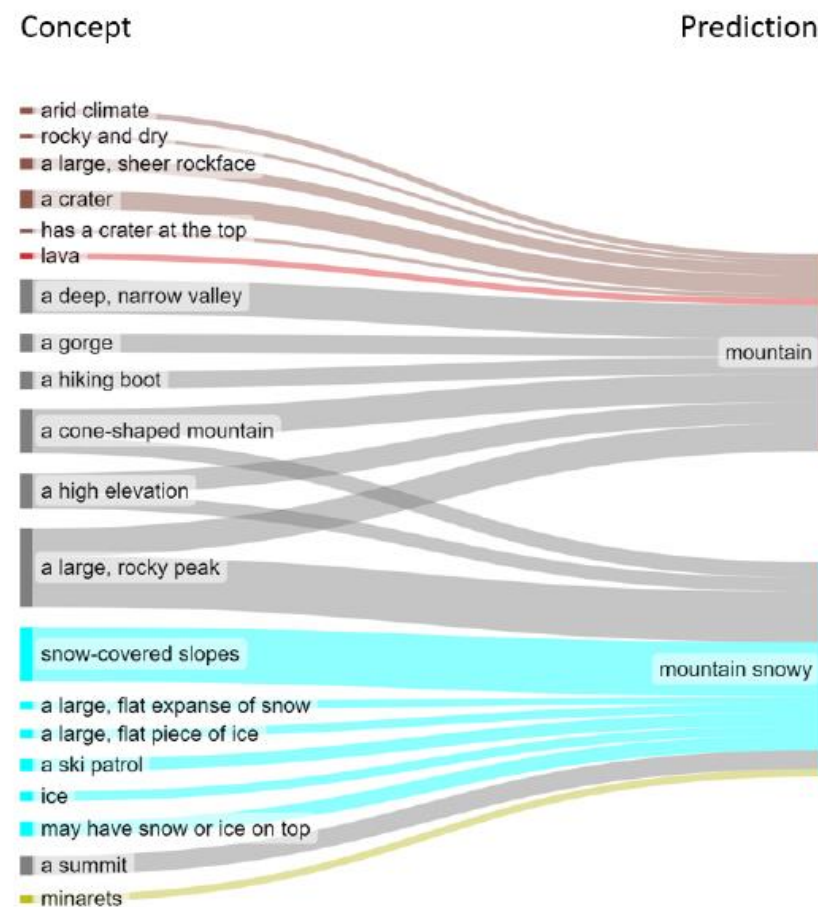
Model	Sparse final layer	Dataset				
		CIFAR10	CIFAR100	CUB200	Places365	ImageNet
Standard	No	88.80%*	70.10%*	76.70%	48.56%	76.13%
Standard (sparse)	Yes	82.96%	58.34%	75.96%	38.46%	74.35%
P-CBM	Yes	70.50%*	43.20%*	59.60%*	N/A	N/A
P-CBM (CLIP)	Yes	84.50%*	56.00%*	N/A	N/A	N/A
Label-free CBM (Ours)	Yes	86.40% $\pm 0.06\%$	65.13% $\pm 0.12\%$	74.31% $\pm 0.29\%$	43.68% $\pm 0.10\%$	71.95% $\pm 0.05\%$

Result

ImageNet CBM
Orange vs Lemon

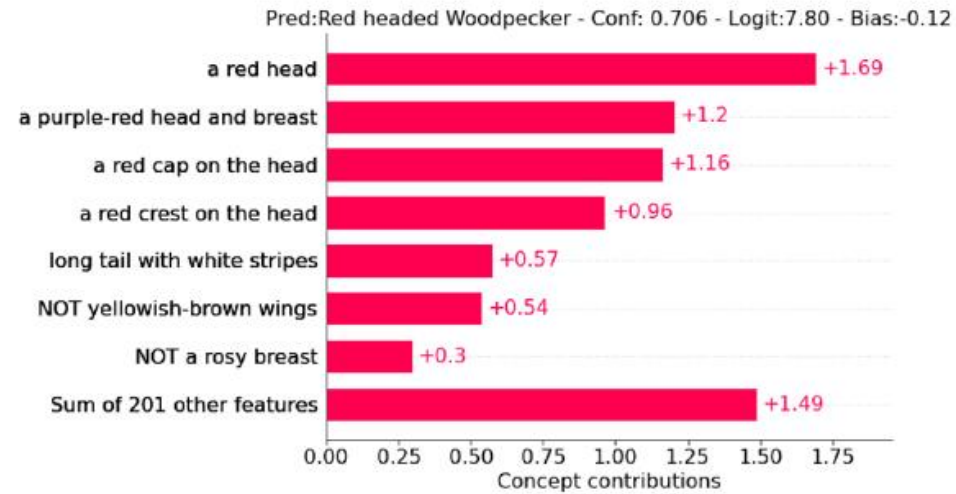


Places365 CBM
Mountain vs Mountain Snowy



Result

CUB

Places
365