Catch Your Concepts: A Flexible Concept Locator for Interpretable Visual Recognition

Qiyang Wan, Ruiping Wang, Chengzhi Gao, Xilin Chen

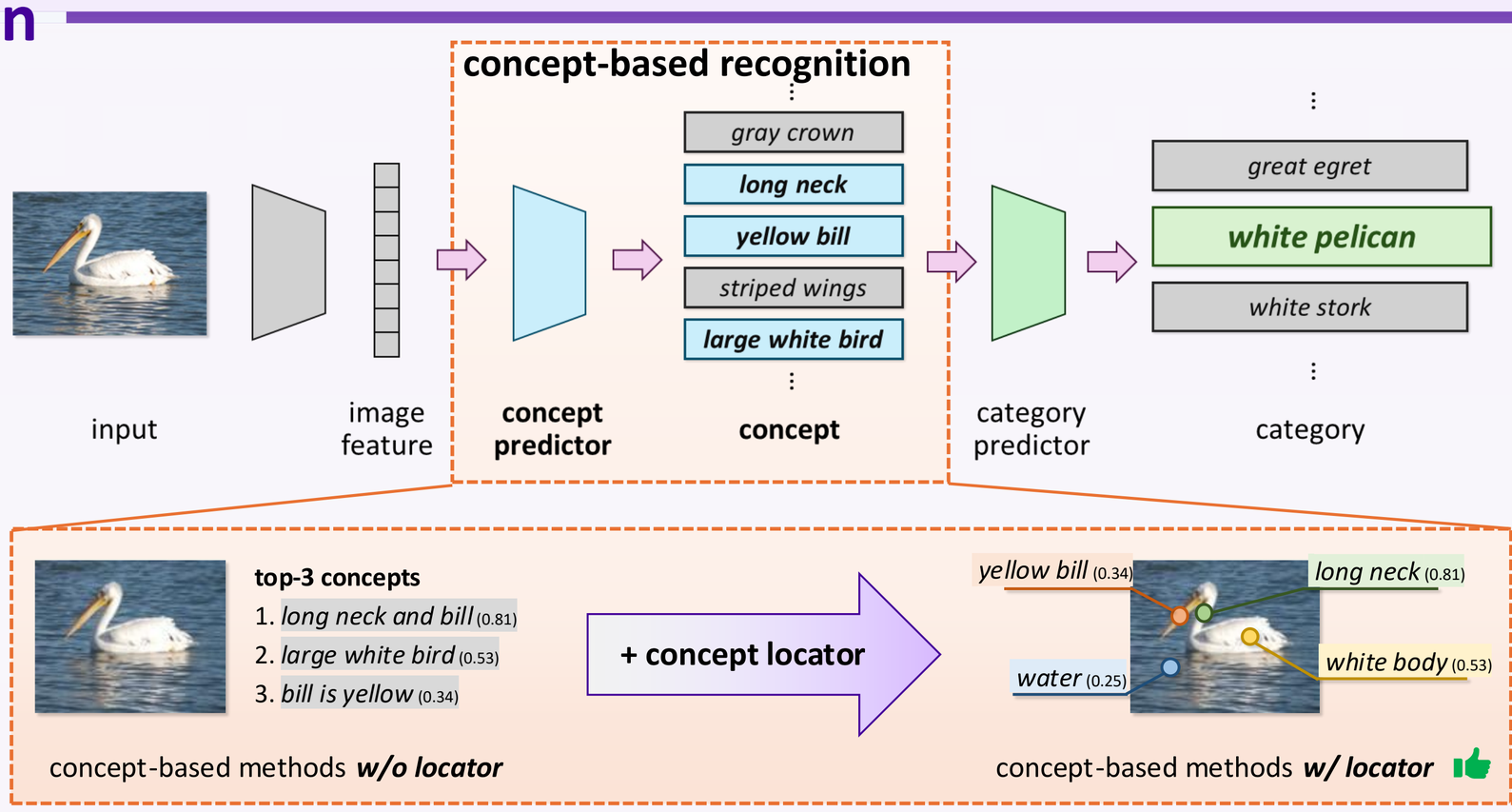
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

{qiyang.wan, chengzhi.gao}@vipl.ict.ac.cn, {wangruiping, xlchen}@ict.ac.cn

Code is available at *https://github.com/VIPL-VSU/Loc-CBM*

Background & Motivation

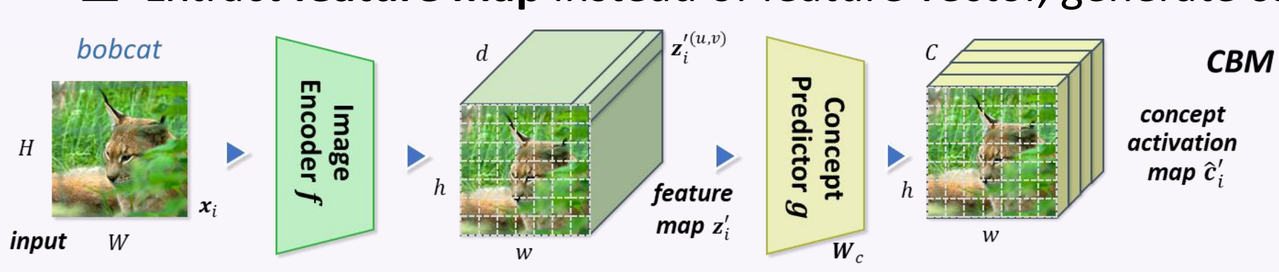
- **Background**
  - Black-box models are with limited interpretability
  - **Concept-based methods** explain predictions via understandable concepts, but **lack spatial grounding**
- **Motivation**
  - Adding a **concept locator** links ***what*** and ***where***
  - provide both **semantic** and **localization** explanations

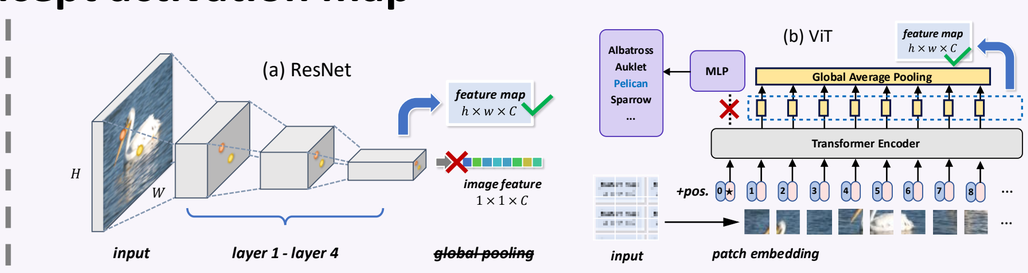


Method

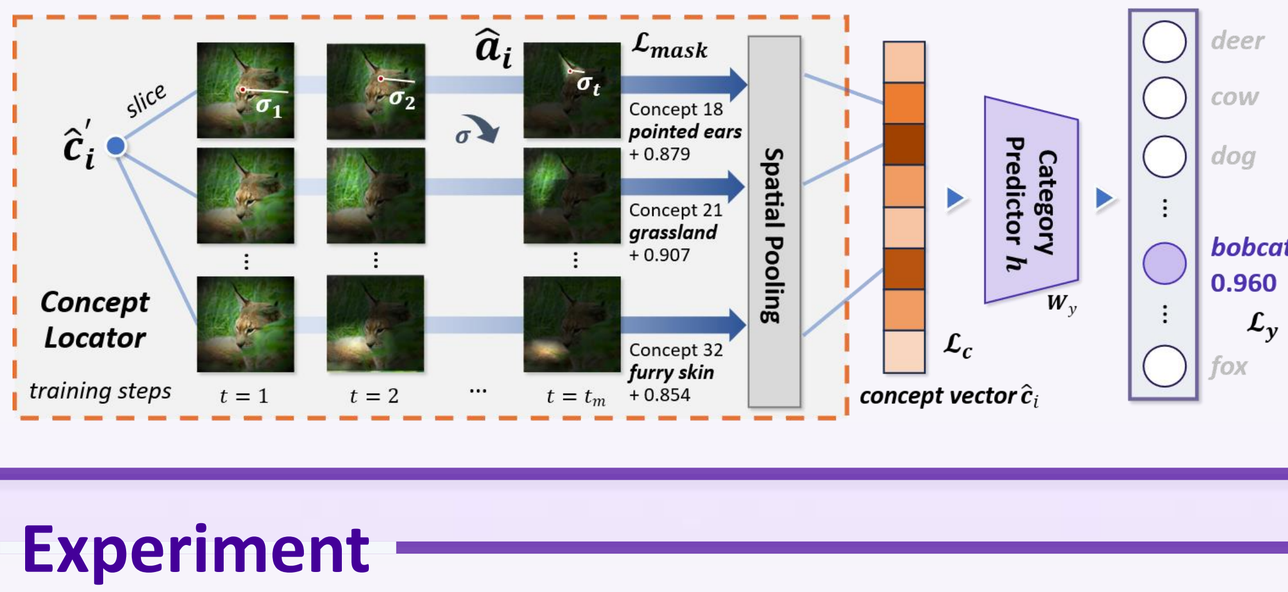
- **Modified concept-based framework**: **preserve spatial information** for localization
  - Extract **feature map** instead of feature vector, generate **concept activation map**

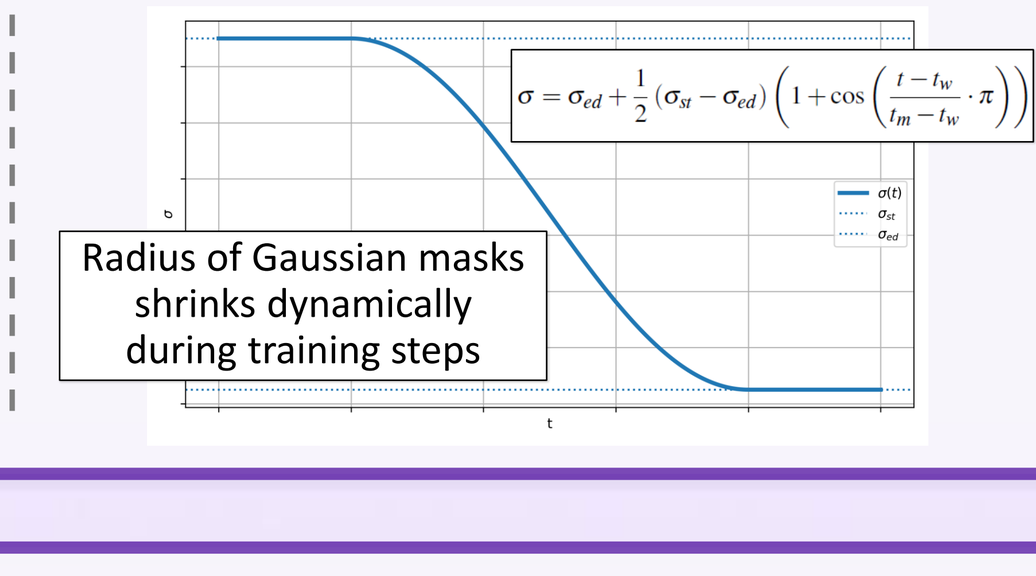
Flexible to various backbone types





- **Concept locator**: constrain concept activation in **concentrated spatial regions** by applying masks
  - Find **activation peaks** and build **Gaussian masks**, apply and learn via **alignment loss** $\mathcal{L}_{mask} = \mathbb{E}[\mathcal{L}_{MSE}(\hat{c}_i', \hat{a}_i)]$





Experiment

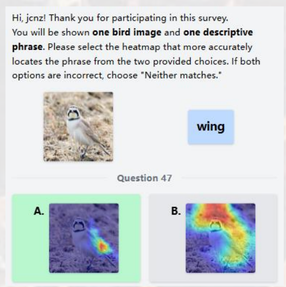
- **Recognition Performance**
  - Loc-CBM outperforms SOTAs on multiple datasets

Method	CIF-10	CIF-100	CUB/18	IN-1k	Places
Vanilla ResNet	93.7	74.0	76.7	76.4	53.4
LFCBM [ICLR'23]	86.4	65.1	74.3	67.5	49.0
LaBo [CVPR'23]	87.9	69.1	71.8	68.9	N/A
Res-CBM [CVPR'24]	88.0	67.9	62.4	N/A	N/A
DN-CBM [ECCV'24]	87.6	67.5	N/A	72.9	53.5
VLG-CBM [NeurIPS'24]	88.6	66.5	75.8	62.7	N/A
Loc-CBM (ours)	**92.9**	**73.0**	**76.1**	**73.6**	**54.2**

Method	CUB/50
Vanilla ResNet	84.5
ProtoPNet [NeurIPS'19]	79.4
ProtoTree [CVPR'21]	82.2
ProtoPool [ECCV'22]	85.5
PIP-Net [CVPR'23]	82.0
SPANet [WACV'24]	81.7
Loc-CBM (ours)	85.1

- **Localization Performance**
  - Loc-CBM localizes concepts more accurately

Localization Metric	Base CBM	Loc-CBM
Keypoint Alignment Error ↓	6.892	**0.959**
Concept Activation L1-norm ↓	3.902	**2.604**
User Satisfaction ↑	0.153	**0.317**



User Study Interface

- **Visualization**
  - activation maps are much **cleaner** and **more precise** than before
  - localizes meaningful concepts **across datasets**