國立陽明交通大學
NATIONAL YANG MING CHIAO TUNG UNIVERSITY

Paper Tag: 3-119

NVIDIA

MCPNet: An Interpretable Classifier via Multi-Level Concept Prototypes



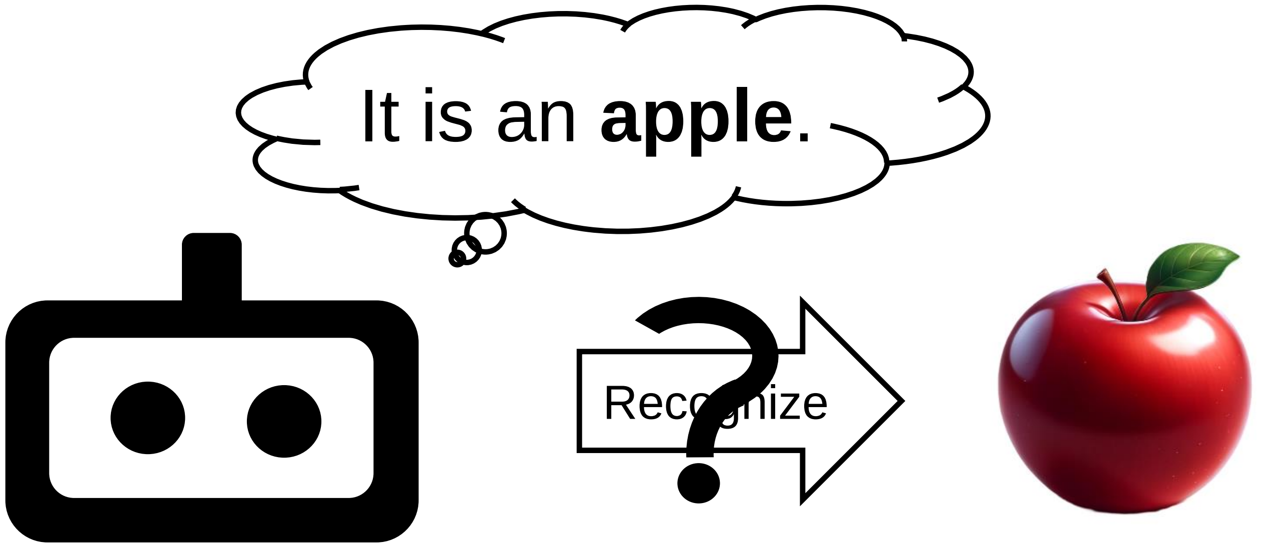
Bor-Shiun Wang   Chien-Yi Wang   Wei-Chen Chiu

Enriched Vision Applications Laboratory

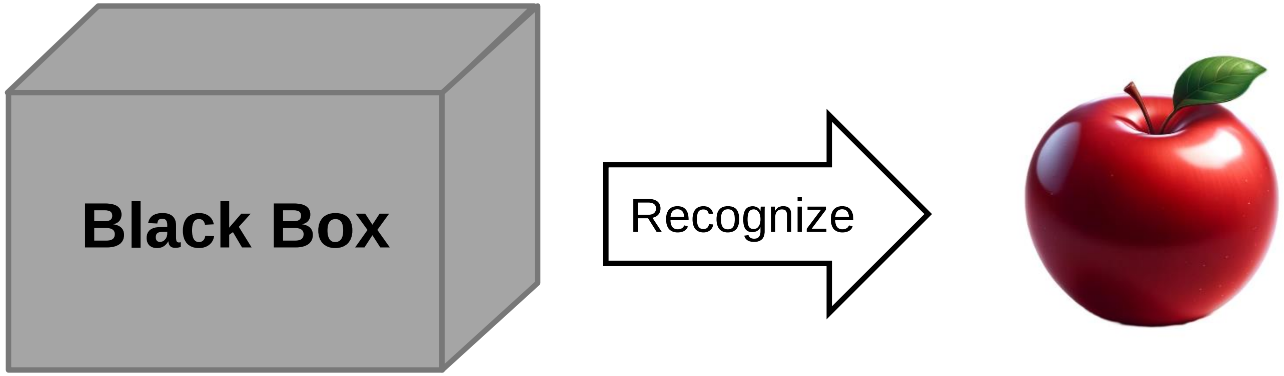
SEATTLE, WA
CVPR
JUNE 17-21, 2024

What is Explainable AI (XAI)?

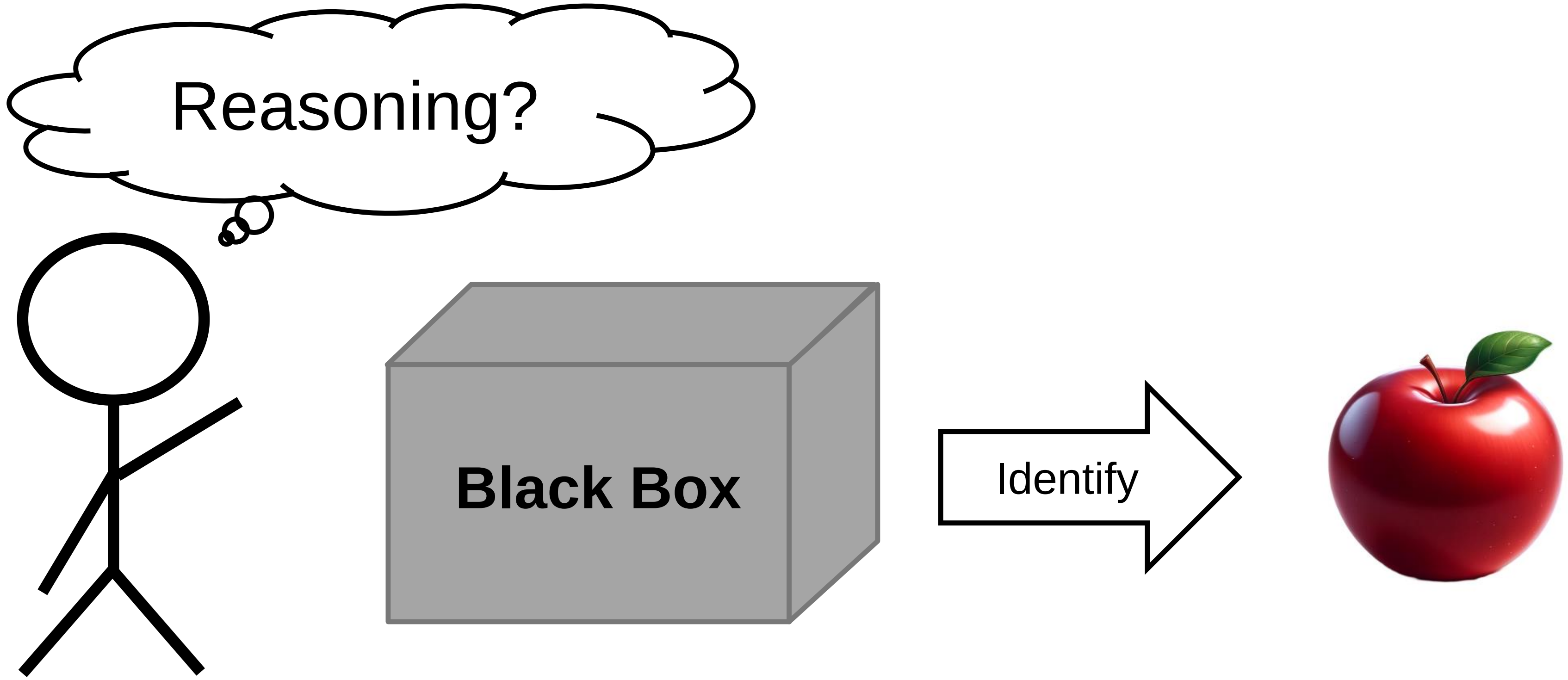
What is Explainable AI (XAI)?



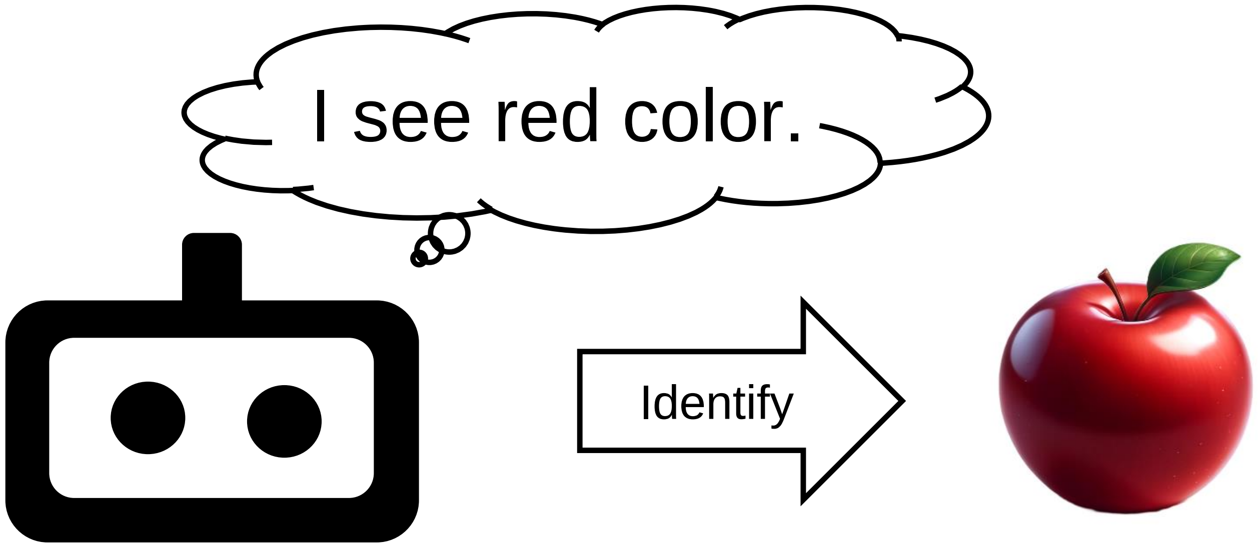
What is Explainable AI (XAI)?



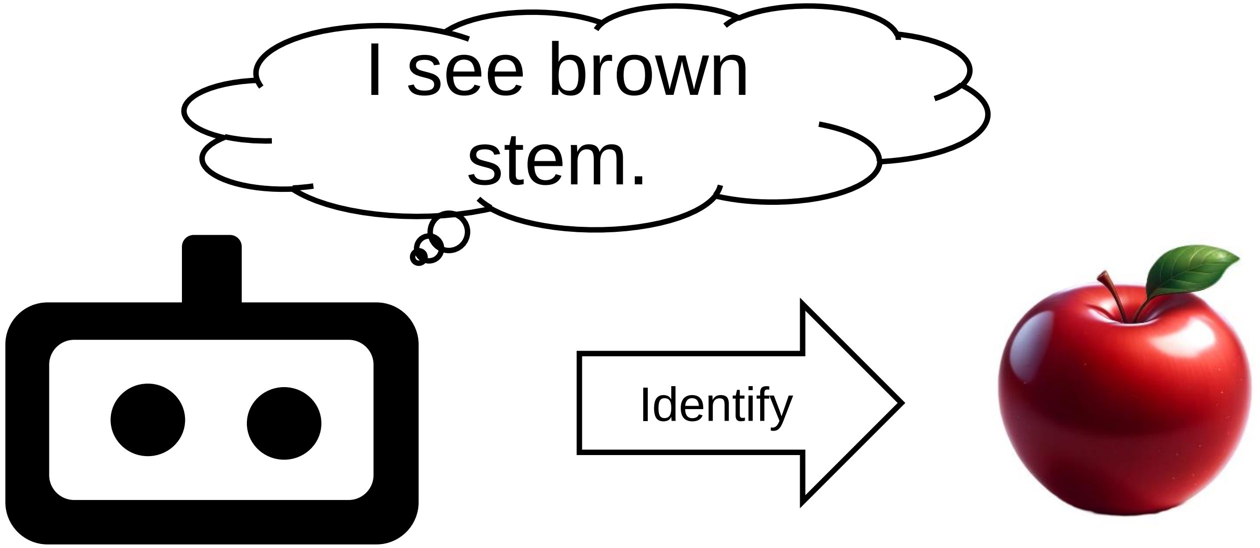
What is Explainable AI (XAI)?



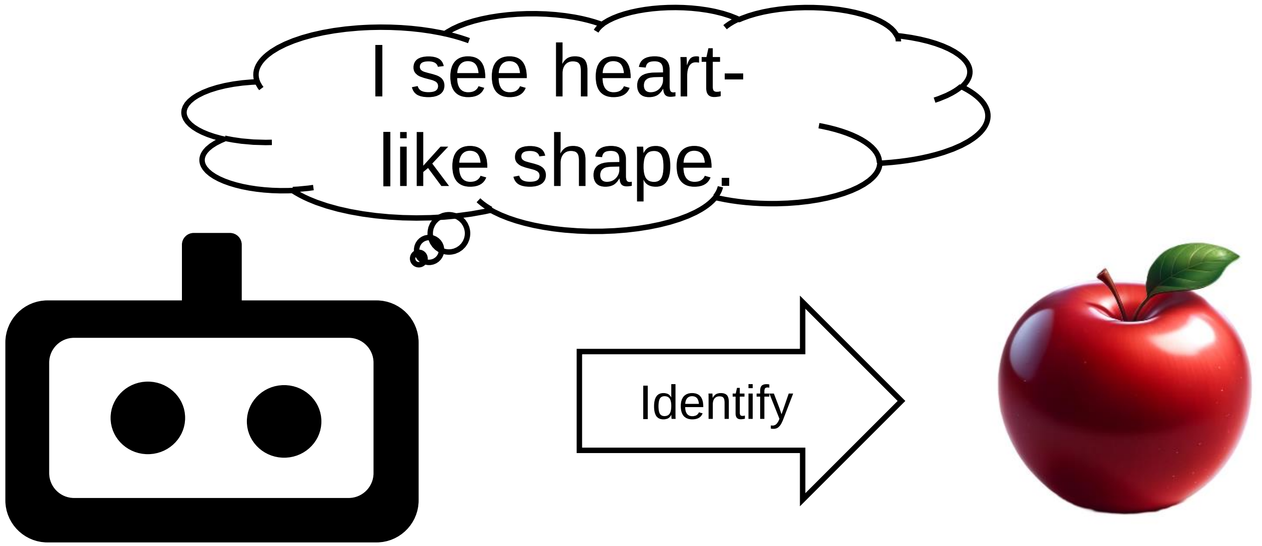
What is Explainable AI (XAI)?



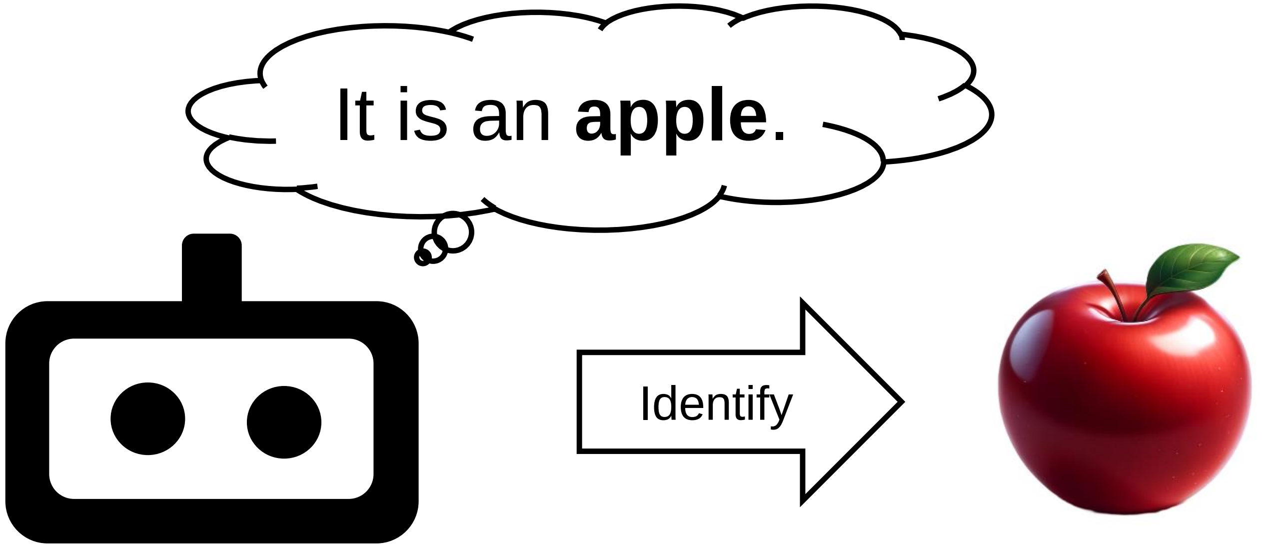
What is Explainable AI (XAI)?



What is Explainable AI (XAI)?



What is Explainable AI (XAI)?



Types of Explainable AI Methods

- Two types of explainable AI methods
  - Post-hoc method
    - Explainable the model after the training
    - Inconsistent to prediction
  - Inherent method
    - Model is designed with interpretability built into their structure
    - Performance trade-offs between accuracy and interpretability

Types of Explainable AI Methods

- Two types of explainable AI methods
  - Post-hoc method
    - Explainable the model after the training
    - Inconsistent to prediction
  - Inherent method
    - Model is designed with interpretability built into their structure
    - Performance trade-offs between accuracy and interpretability

Types of Explainable AI Methods

- Two types of explainable AI methods
  - Post-hoc method
    - Explainable the model after the training
    - Inconsistent to prediction
  - Inherent method
    - Model is designed with interpretability built into their structure
    - Performance trade-offs between accuracy and interpretability

Motivation

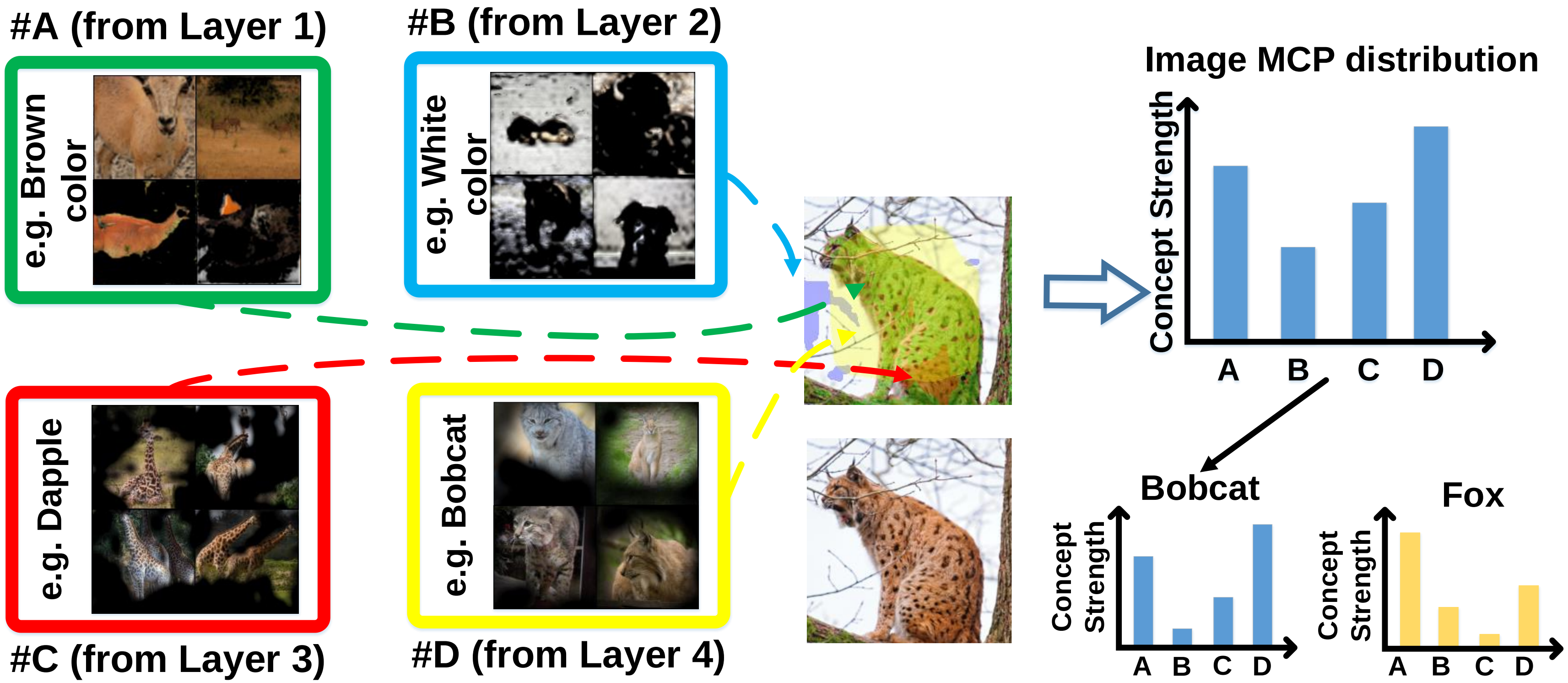
- Providing low and mid-level explanations:
  - A more comprehensive aspect to unveil the model
  - Seamlessly integrated into various model (CNN-based)

	MCPNet(Ours)	ProtoPNet [1,2,3]	Concept Bottleneck [4]	TCAV [5]	CRAFT [6]
Explanation Type	Inherent	Inherent	Inherent	Post-hoc	Post-hoc
Explanation Scale	Multi-Level	Single-Level	Single-Level	Single-Level	Single-Level
w/o Concept Labels	✓	✓	✗	✗	✓
w/o Modifying Models	✓	✗	✗	✓	✓

[1] CHEN, Chaofan, et al. This looks like that: deep learning for interpretable image recognition.
[2] DONNELLY, Jon, et al. Deformable protopnet: An interpretable image classifier using deformable prototypes.
[3] NAUTA, Meike, et al. Pip-net: Patch-based intuitive prototypes for interpretable image classification.
[4] KOH, Pang Wei, et al. Concept bottleneck models. In: *International conference on machine learning*.
[5] KIM, Been, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).
[6] FEL, Thomas, et al. Craft: Concept recursive activation factorization for explainability.

Target

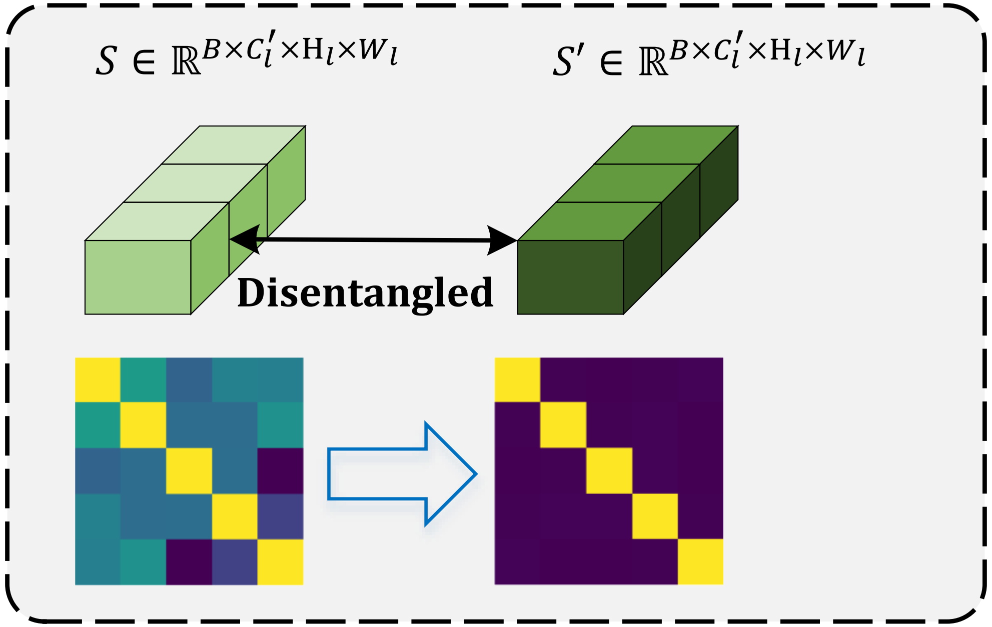
- An <u>inherently hierarchical explanation method</u> to unveil the model
  - Providing multi-scale explanations
  - Without compromising the performance
  - Seamlessly integrate with various backbone (CNN-based)



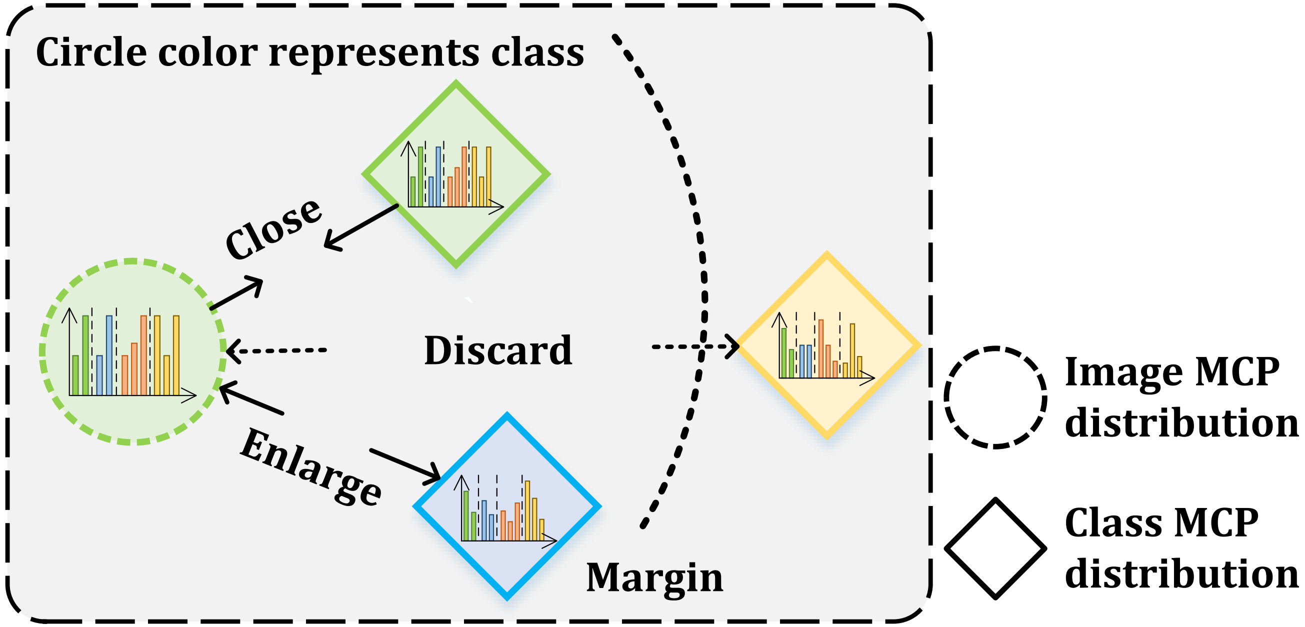
Proposed Method

- Proposed constraints
  - Centered Kernel Alignment (CKA) loss
  - Class-aware Concept Distribution (CCD) loss

Centered Kernel Alignment loss

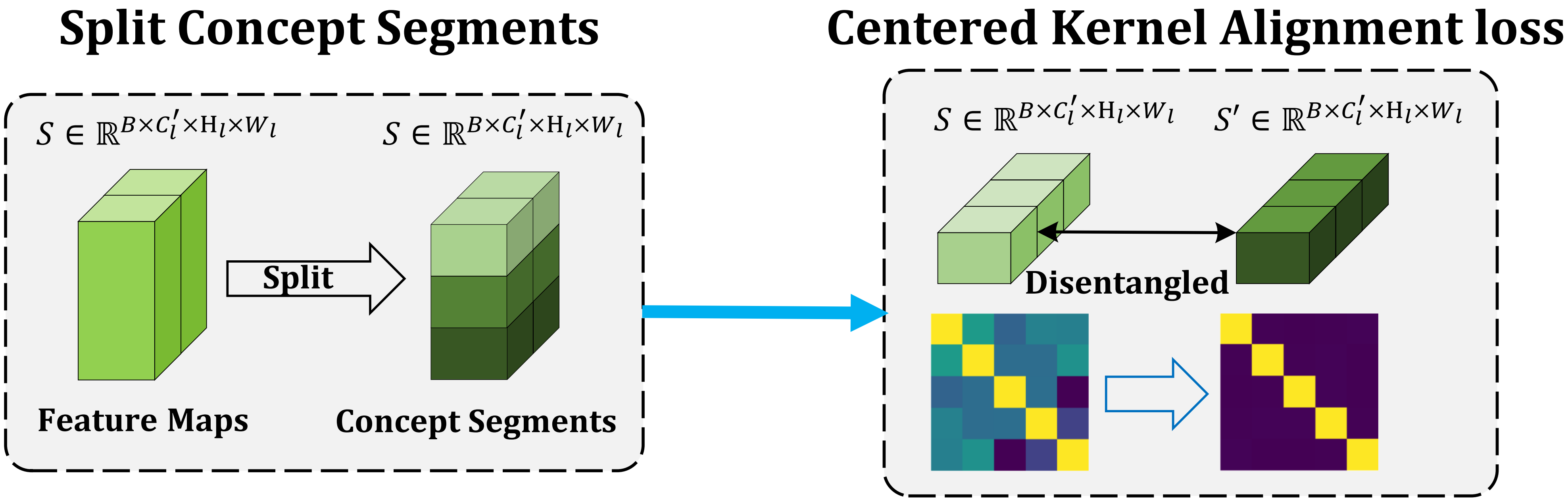


Class-aware Concept Distribution loss



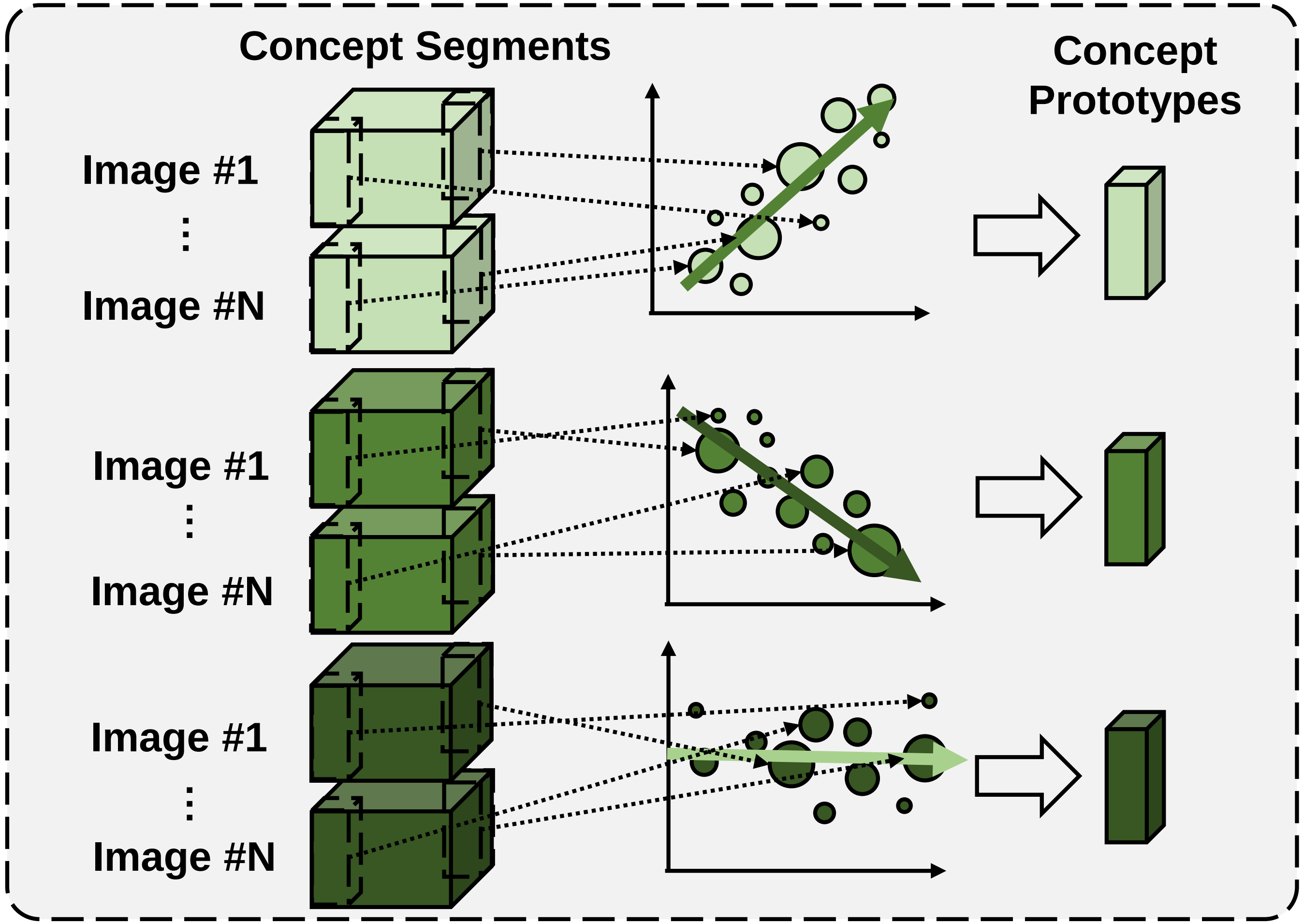
Proposed Method

- Centered Kernel Alignment (CKA) loss
  - Disentangling segment semantics



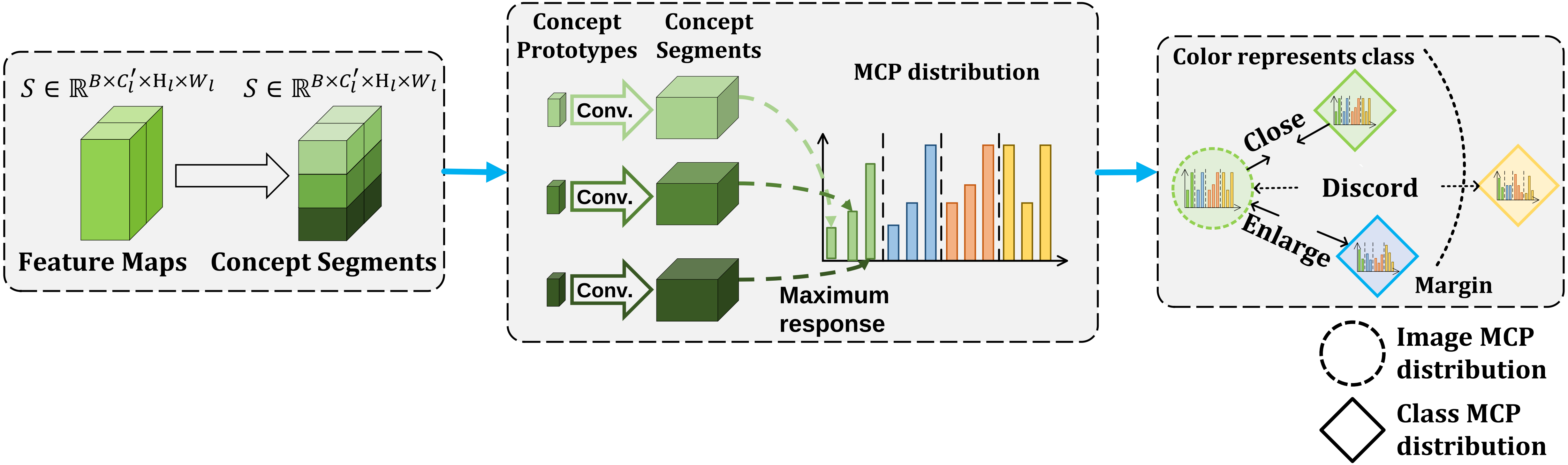
Proposed Method

- Concept prototypes extraction



Proposed Method

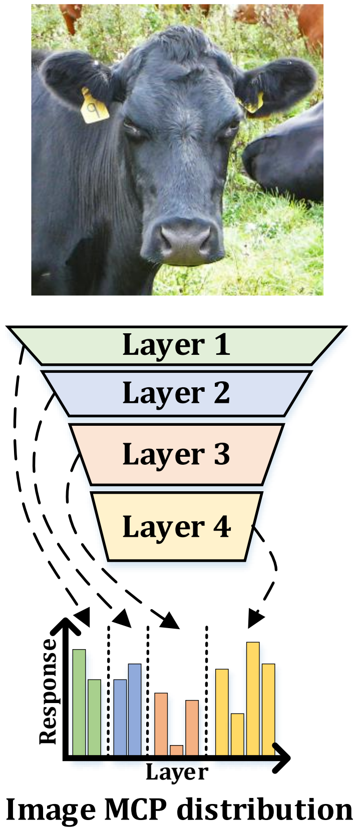
- Class-aware Concept Distribution (CCD) loss
  - Classifying via Multi-level Concept Prototypes distributions (MCP distribution)



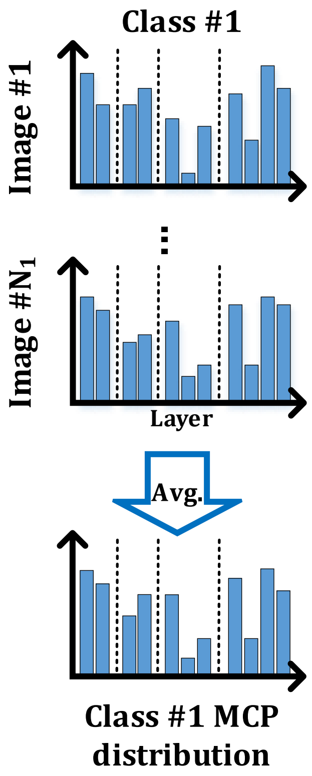
Proposed Method

- Our new classify paradigm

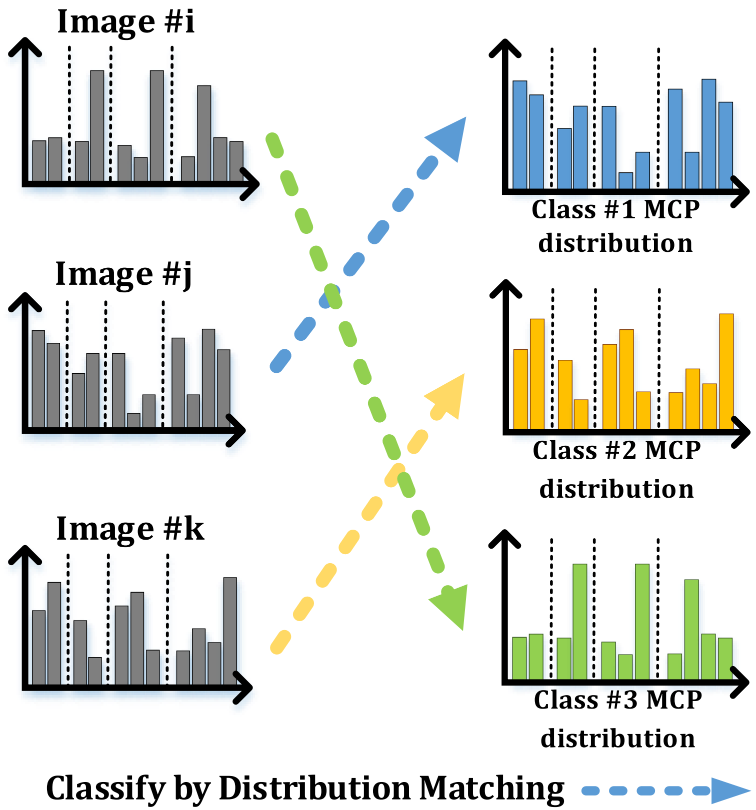
For each image:



For each class:



Classify images:



Experiments – Quantitative Results

- Main quantitative results:

Backbone	Methods	Explanation	Accuracy AWA2	Accuracy Caltech101	Accuracy CUB_200_2011
ResNet50	Baseline	N/A	94.92%	94.21%	77.94%
	ProtoTree [14]	Single-Scale	90.60%	72.19%	18.00%†
	Deformable ProtoPNet [2]	Single-Scale	85.51%	93.88%	73.42%†
	ST-ProtoPNet [28]	Single-Scale	93.76%	95.95%	76.34%†
	PIP-Net [15]	Single-Scale	85.99%	87.86%	70.99%†
	MCPNet (Ours)	Multi-Scale	93.92%	93.88%	80.15%
Inception V3	Baseline	N/A	95.47%	96.42%	79.43%
	ProtoTree [14]	Single-Scale	92.29%	86.02%	13.03%
	Deformable ProtoPNet [2]	Single-Scale	92.68%	97.22%	72.99%
	ST-ProtoPNet [28]	Single-Scale	93.60%	96.99%	75.25%
	PIP-Net [15]	Single-Scale	43.82%	45.04%	6.76%
	MCPNet (Ours)	Multi-Scale	94.62%	95.76%	78.94%
ConvNeXt-tiny	Baseline	N/A	96.55%	96.56%	84.55%
	ProtoTree [14]	Single-Scale	94.00%	78.82%	21.57%
	Deformable ProtoPNet [2]	Single-Scale	91.94%	93.65%	35.05%
	ST-ProtoPNet [28]	Single-Scale	94.22%	97.17%	81.84%
	PIP-Net [15]	Single-Scale	93.80%	96.61%	82.74%
	MCPNet (Ours)	Multi-Scale	95.61%	95.95%	83.45%

Experiments – 5-shot Classification

- 5-shot for unseen class images classification:

Dataset	Method	Accuracy
AWA2	Baseline	60.55%
	ProtoTree [14]	33.68%
	Deformable ProtoPNet [2]	19.71%
	ST-ProtoPNet [28]	30.15%
	PIP-Net [15]	26.17%
	MCPNet (Ours)	**73.79%**

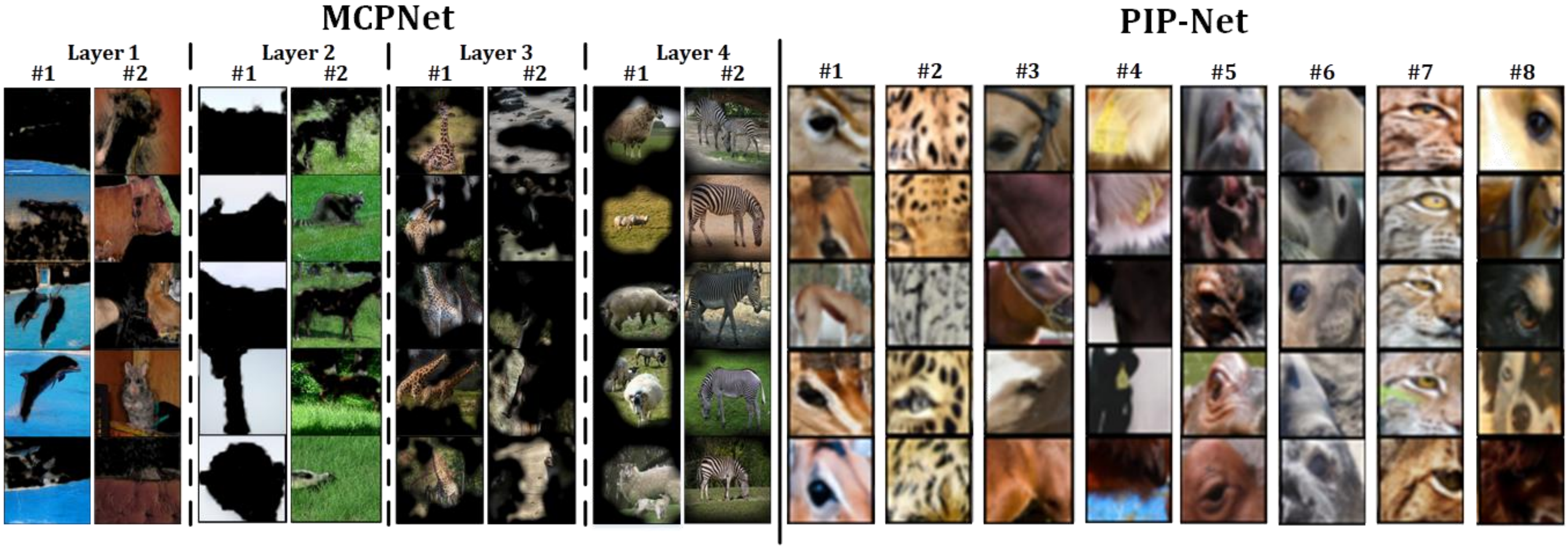
Experiments – Ablation Study

- The effect of different number of channel per segment:

Dataset	Channel	Accuracy
AWA2	32	93.92%
	16	93.95%
	8	93.58%
Caltech101	32	93.88%
	16	93.79%
	8	93.51%
CUB_200_2011	32	80.15%
	16	80.19%
	8	81.22%

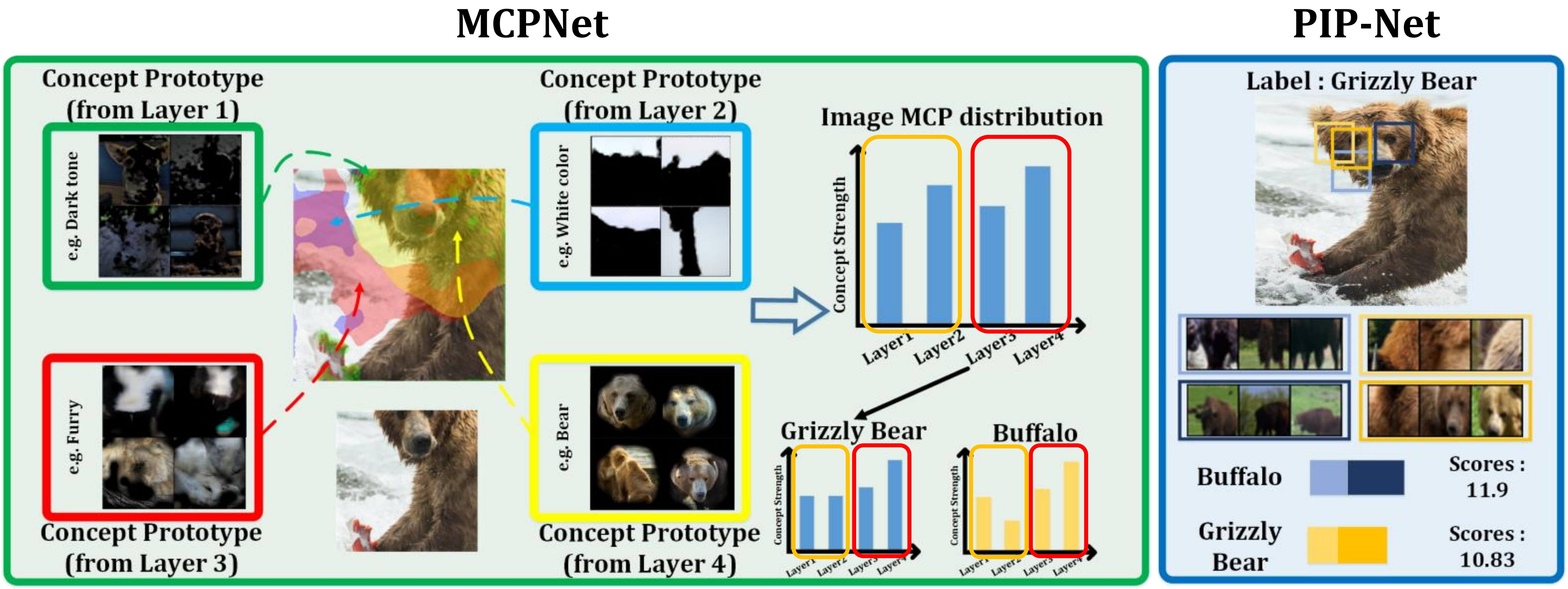
Experiments – Concept Visualizations

- MCPNet (Ours) provides different scales concepts.
- Previous methods (e.g. PIP-Net) only provides single scale concepts.



NAUTA, Meike, et al. Pip-net: Patch-based intuitive prototypes for interpretable image classification.

Experiments - Explanations



國立陽明交通大學
NATIONAL YANG MING CHIAO TUNG UNIVERSITY

NVIDIA

Paper Tag: 3-119

MCPNet: An Interpretable Classifier via Multi-Level Concept Prototypes


Bor-Shiun Wang


Chien-Yi Wang


Wei-Chen Chiu

Project page:

Enriched Vision Applications Laboratory

SEATTLE, WA

CVPR
JUNE 17-21, 2024