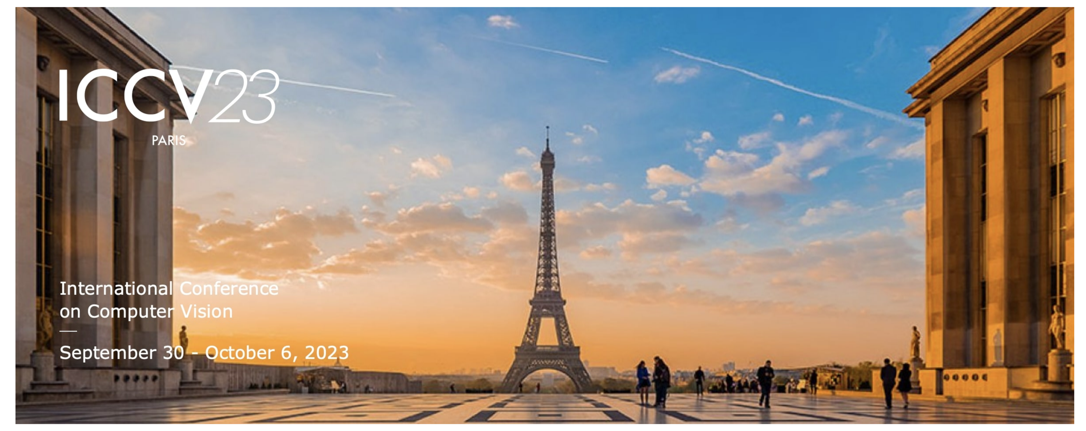
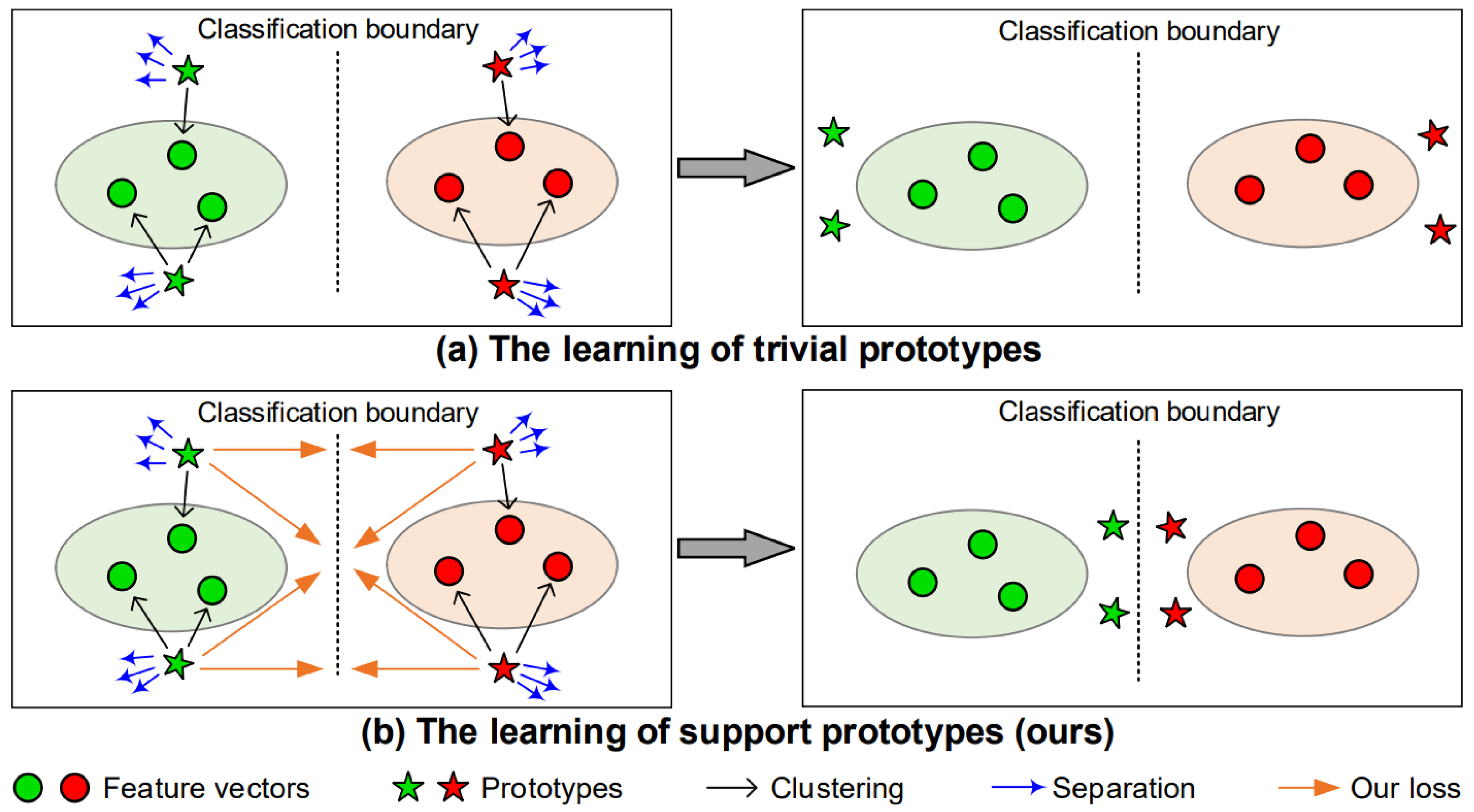


Learning Support and Trivial Prototypes for Interpretable Image Classification

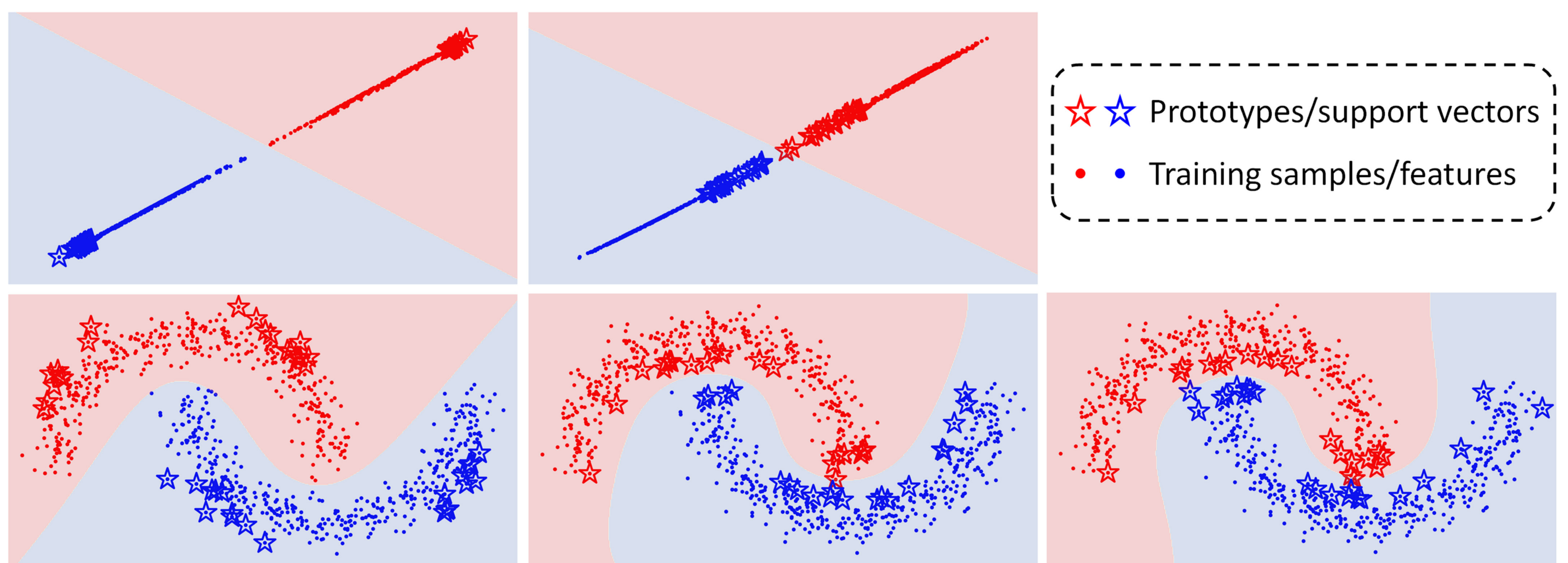
Chong Wang, Yuyuan Liu, Yuanhong Chen, Fengbei Liu, Yu Tian,
Davis J. McCarthy, Helen Frazer, Gustavo Carneiro

Introduction & Motivation

- Interpretability of deep-learning models is highly demanded in high-stakes applications, e.g., disease diagnosis and autonomous driving.
- ProtoPNet achieves similarity-based interpretable classification by measuring how strongly parts of a test image look like the training prototypes.
- ProtoPNet tends to learn trivial prototypes, due to the co-effects of clustering and separation training losses.



- We make an analogy between the prototype learning from ProtoPNet and support vector learning from SVM, and propose to learn support prototypes that benefit classification accuracy and interpretability.

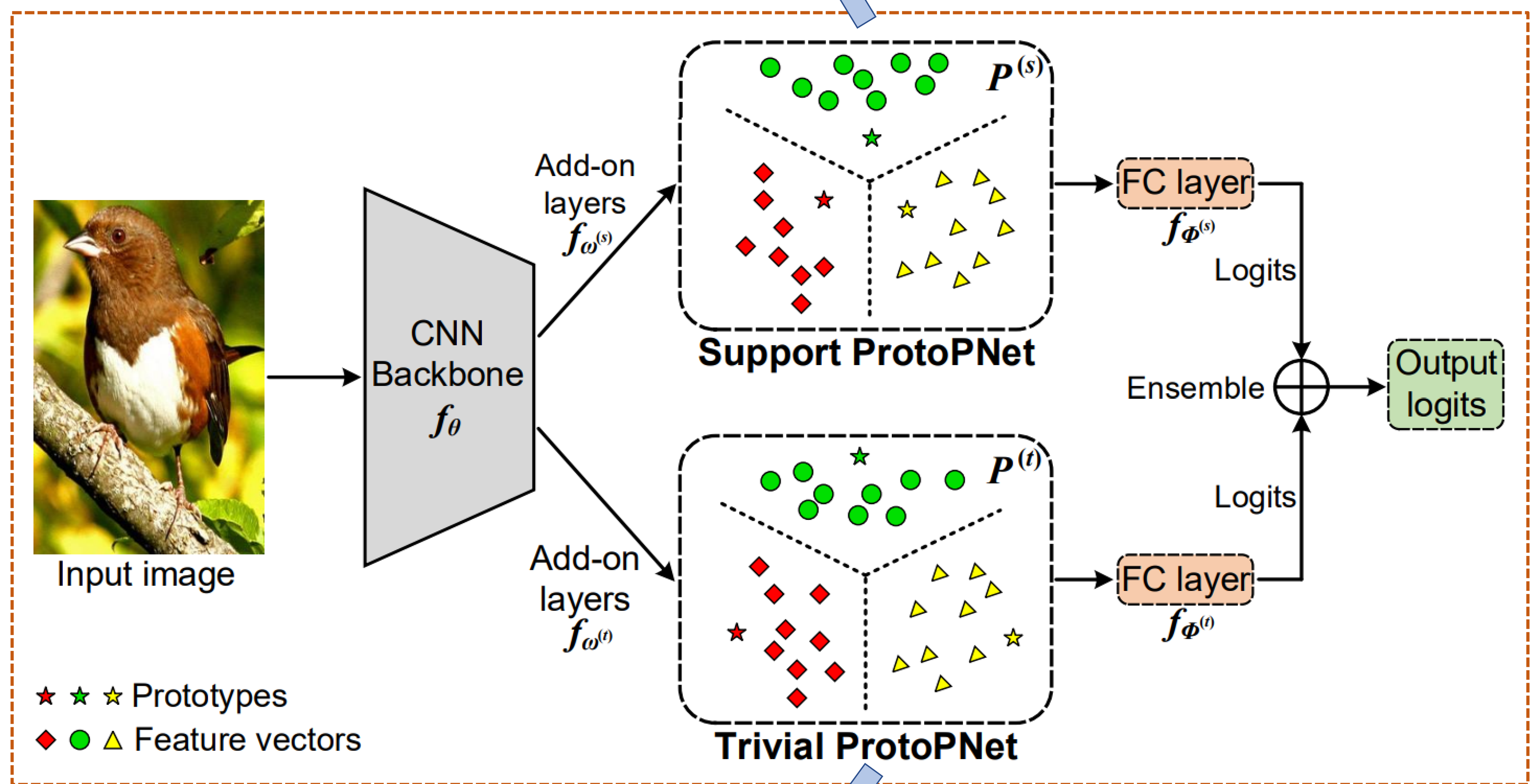


(a) Trivial Prototypes (b) Support Prototypes (c) Support vectors from SVM

Method

- A support ProtoPNet branch to utilize support prototypes, capturing hard-to-learn visual patterns. (closeness)
- A trivial ProtoPNet branch to employ trivial prototypes, capturing easy-to-learn visual features. (discrimination)
- ST-ProtoPNet: ensemble classification interpretation by the two complementary sets of prototypes.

$$\ell_{cls}(\mathcal{P}^{(s)}) = \sum_{c_1=1}^{C-1} \sum_{c_2=c_1+1}^{C} \min_{\mathbf{p}_m \in \mathcal{P}_{c_1}, \mathbf{p}_n \in \mathcal{P}_{c_2}} \mathbf{p}_m^{\mathrm{T}} \mathbf{p}_n$$



$$\ell_{dsc}(\mathcal{P}^{(t)}) = \sum_{c_1=1}^{C-1} \sum_{c_2=c_1+1}^{C} \max_{\mathbf{p}_m \in \mathcal{P}_{c_1}, \mathbf{p}_n \in \mathcal{P}_{c_2}} \mathbf{p}_m^{\mathrm{T}} \mathbf{p}_n$$

Dataset

- Fine-grained image recognition tasks on CUB-200-2011, Stanford Cars, and Stanford Dogs.
- Evaluation metrics:

  Classification: top-1 accuracy

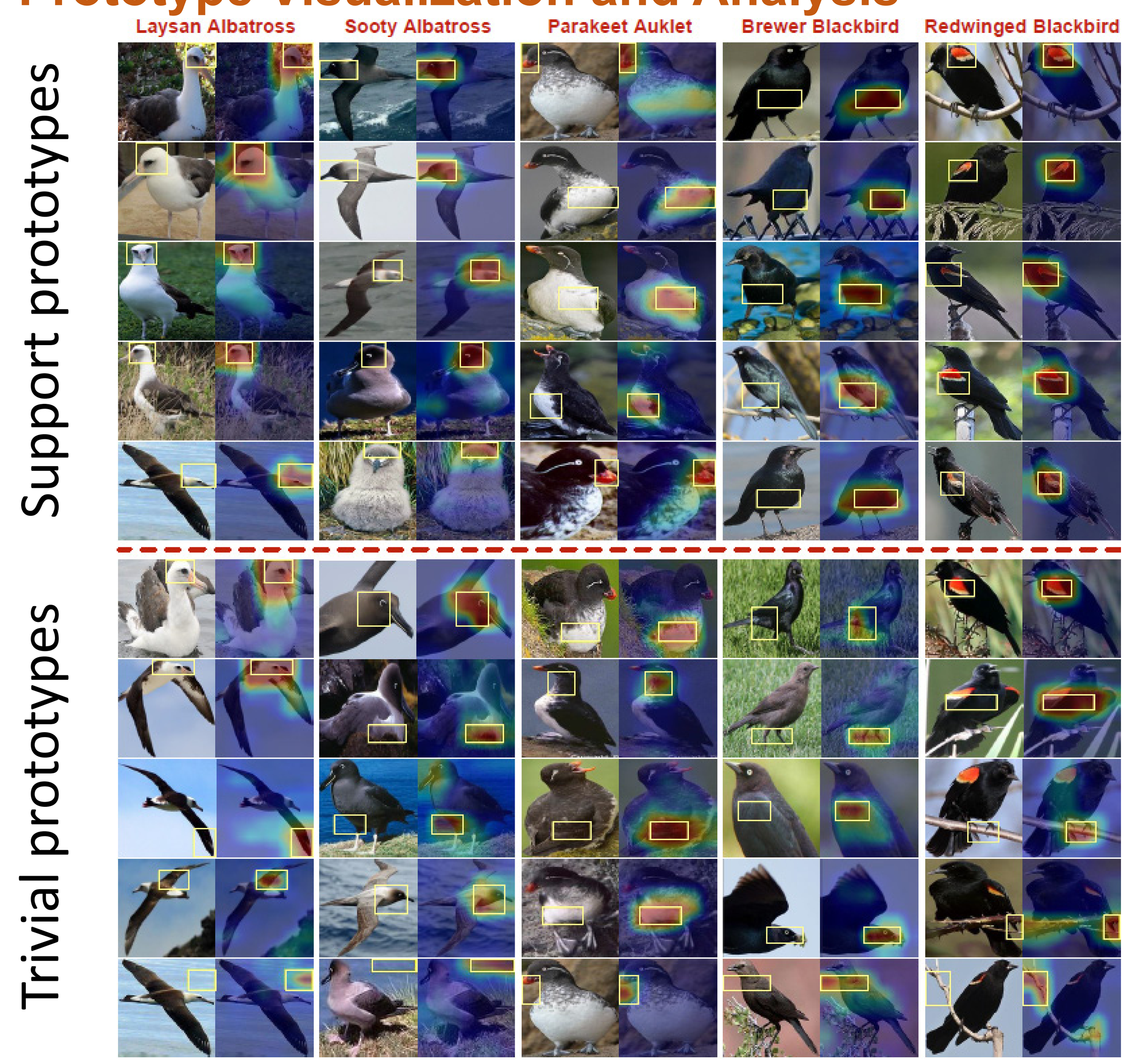
  Interpretability: CH, OIRR, IoU, and DAUC

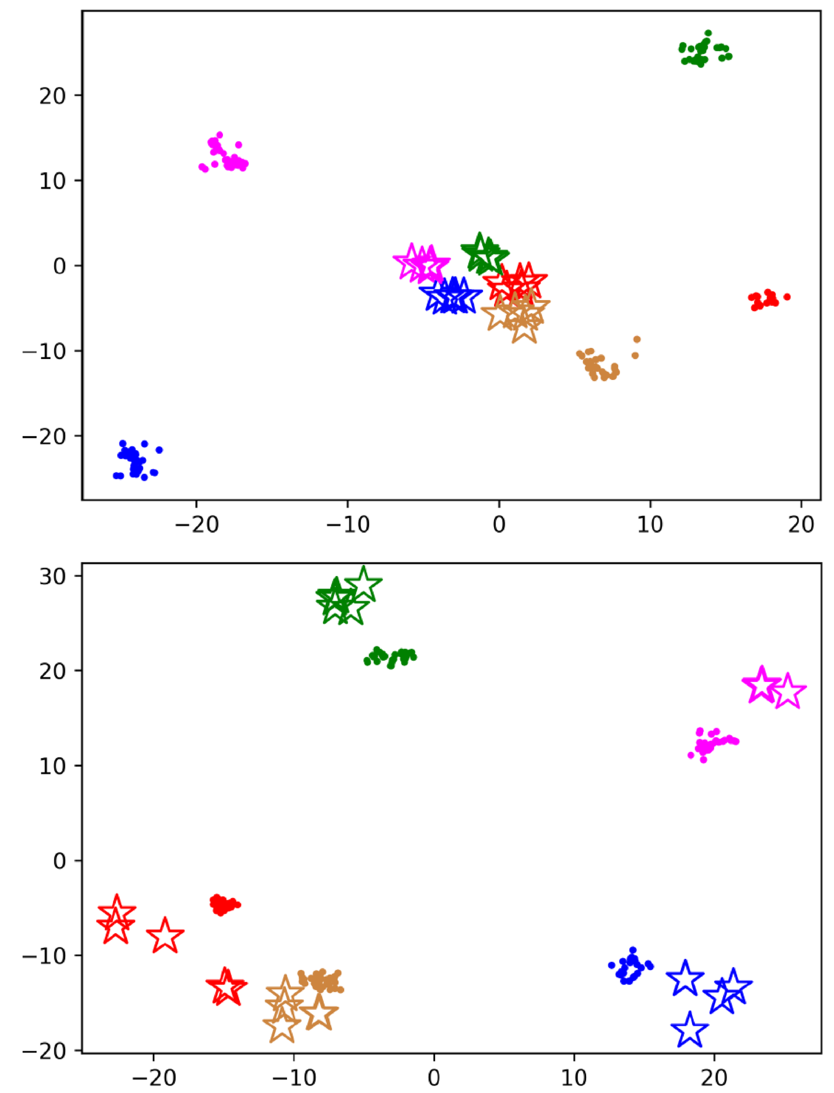
Experimental Results

Classification Accuracy

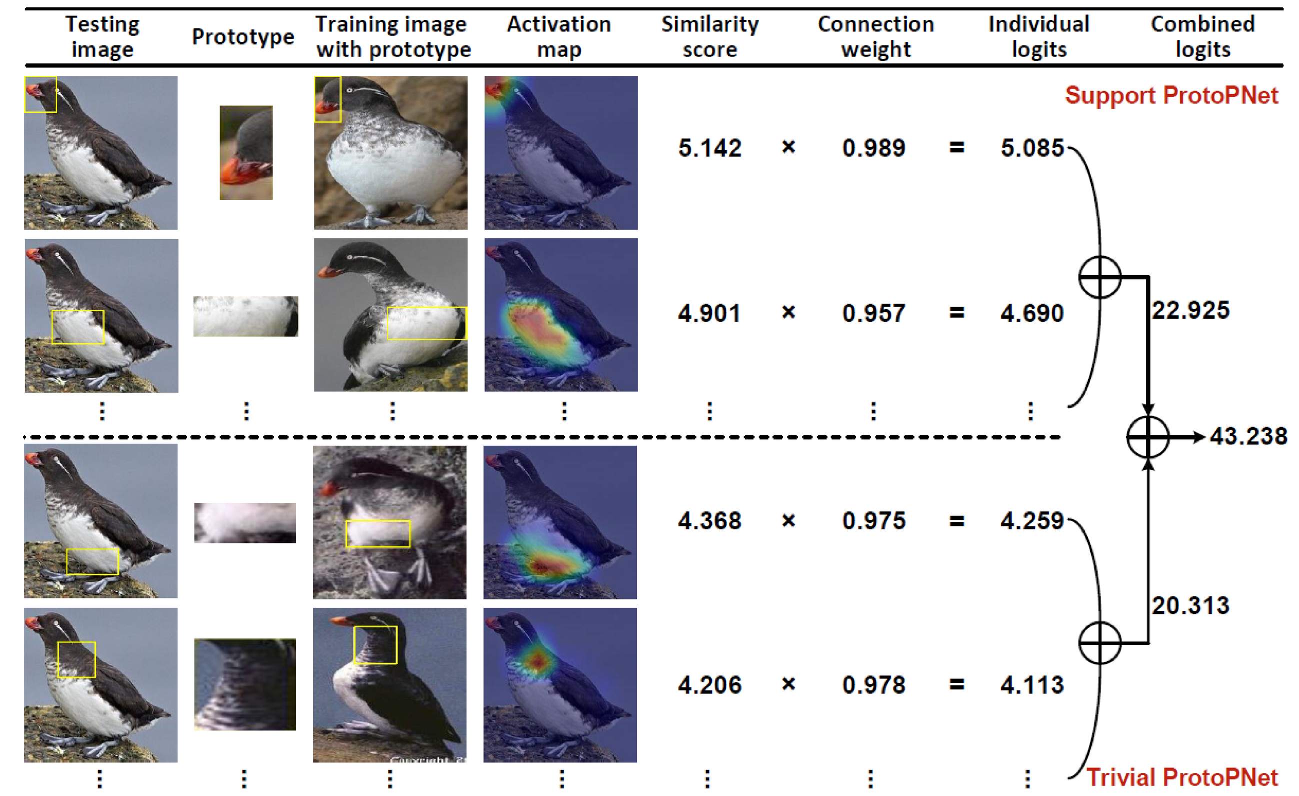
Method	CUB						Cars					
	VGG16	VGG19	ResNet34	ResNet152	Dense121	Dense161	VGG16	VGG19	ResNet34	ResNet152	Dense121	Dense161
Baseline	73.3 ± 0.2	74.7 ± 0.4	82.2 ± 0.3	80.8 ± 0.4	81.8 ± 0.1	82.1 ± 0.2	87.3 ± 0.4	88.5 ± 0.3	**92.6 ± 0.3**	**92.8 ± 0.4**	92.0 ± 0.3	92.5 ± 0.3
ProtoPNet [4]	77.2 ± 0.2	77.6 ± 0.2	78.6 ± 0.1	79.2 ± 0.3	79.0 ± 0.2	80.8 ± 0.3	88.3 ± 0.2	89.4 ± 0.2	88.8 ± 0.1	88.5 ± 0.3	87.7 ± 0.1	89.5 ± 0.2
TesNet [53]	81.3 ± 0.2	81.4 ± 0.1	82.8 ± 0.1	82.7 ± 0.2	84.8 ± 0.2	84.6 ± 0.3	90.3 ± 0.2	90.6 ± 0.2	90.9 ± 0.2	92.0 ± 0.2	91.9 ± 0.3	92.6 ± 0.3
Trivial ProtoPNet	80.8 ± 0.2	81.2 ± 0.2	82.5 ± 0.2	83.1 ± 0.3	83.9 ± 0.3	84.6 ± 0.3	90.1 ± 0.2	90.7 ± 0.2	91.1 ± 0.2	91.5 ± 0.2	91.4 ± 0.3	92.4 ± 0.3
Support ProtoPNet	81.7 ± 0.2	81.8 ± 0.3	83.0 ± 0.1	83.6 ± 0.2	84.7 ± 0.2	85.2 ± 0.3	90.9 ± 0.2	90.8 ± 0.2	91.0 ± 0.2	91.8 ± 0.2	91.7 ± 0.2	92.7 ± 0.3
ST-ProtoPNet (ours)	**82.9 ± 0.2**	**83.2 ± 0.2**	**83.5 ± 0.1**	**84.1 ± 0.2**	**85.4 ± 0.2**	**86.1 ± 0.2**	**91.1 ± 0.2**	**91.7 ± 0.2**	91.4 ± 0.1	92.0 ± 0.2	**92.3 ± 0.3**	**92.7 ± 0.2**

Prototype Visualization and Analysis





Interpretable Reasoning of ST-ProtoPNet



Measuring Interpretability based on Localisation

Metric	GradCAM [43]	ProtoPNet [4]	TesNet [53]	DefProto [11]	TrvProto	SptProto	ST-Proto
CH (%, ↑)	52.46	48.66	59.38	52.09	63.05	63.87	**66.43**
IoU (%, ↑)	39.91	38.03	36.92	40.77	37.74	**42.04**	41.05
OIRR (%, ↓)	37.01	37.26	38.97	28.68	34.48	28.69	**28.09**
DAUC (%, ↓)	7.01	7.39	5.86	5.99	6.06	5.80	**5.74**

Findings

- Support prototypes tend to only focus on relevant bird parts and share visually similar features among classes.
- Trivial porotypes focus not only on the relevant bird parts but also the background regions.

Acknowledgement:
- BRAIx (MRFAI000090)
- ARC (FT190100525)