PEEB: Part-based Image Classifiers with an Explainable and Editable Language Bottleneck

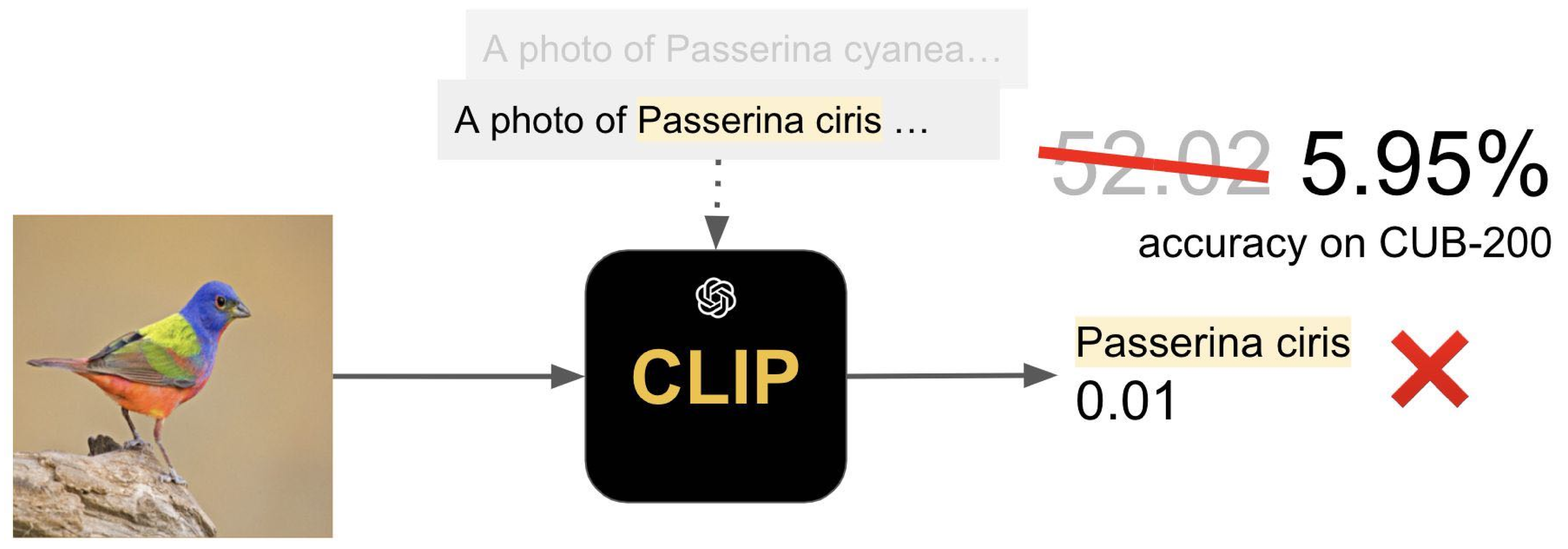
Thang Pham*†, Peijie Chen*†, Tin Nguyen*†, Seunghyun Yoon§, Trung Bui§, Anh Totti Nguyen†

†Auburn University, §Adobe Research

Paper, code & demo: https://github.com/anguyen8/peeb

NAACL 2024

1. Introduction

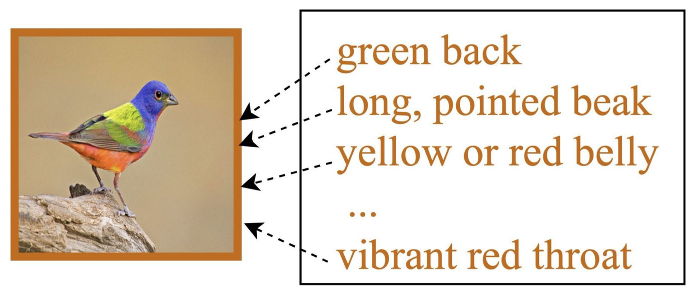


Problems:

1. CLIP relies on known class names.
2. Training required for new, unseen classnames.
3. How text prompts match input images is a black-box process.

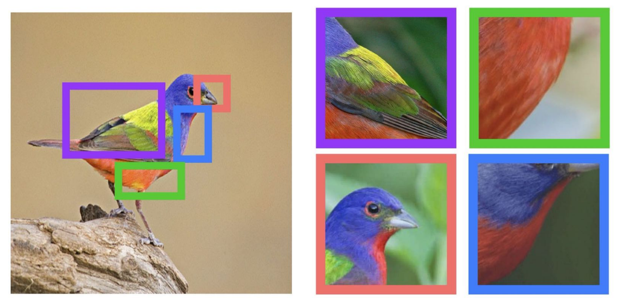
2. Related Work

LaBo (2023), Menon & Vondrick (2023), FuDD (2023), PCBM (2023)

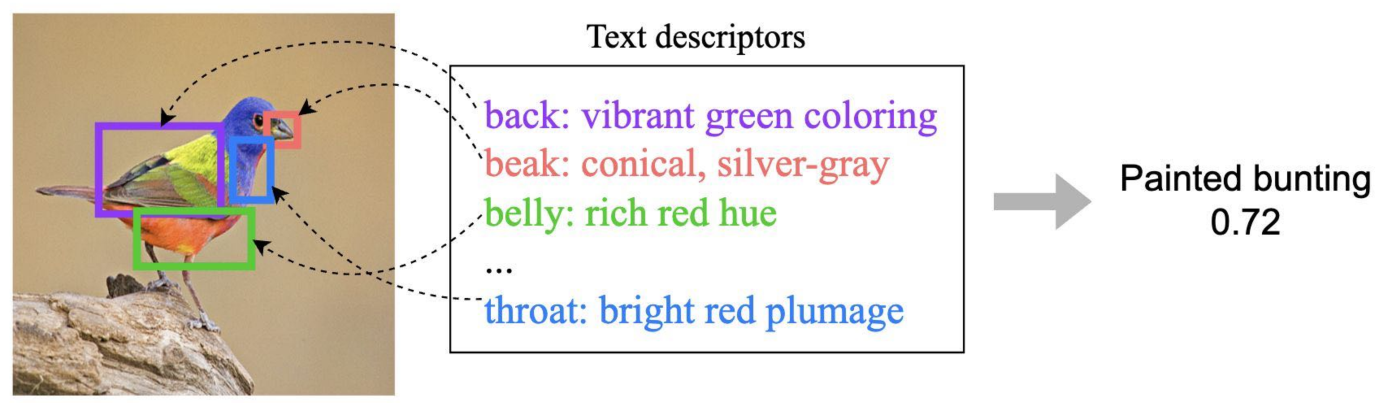


textual concept explanations operating at the image level

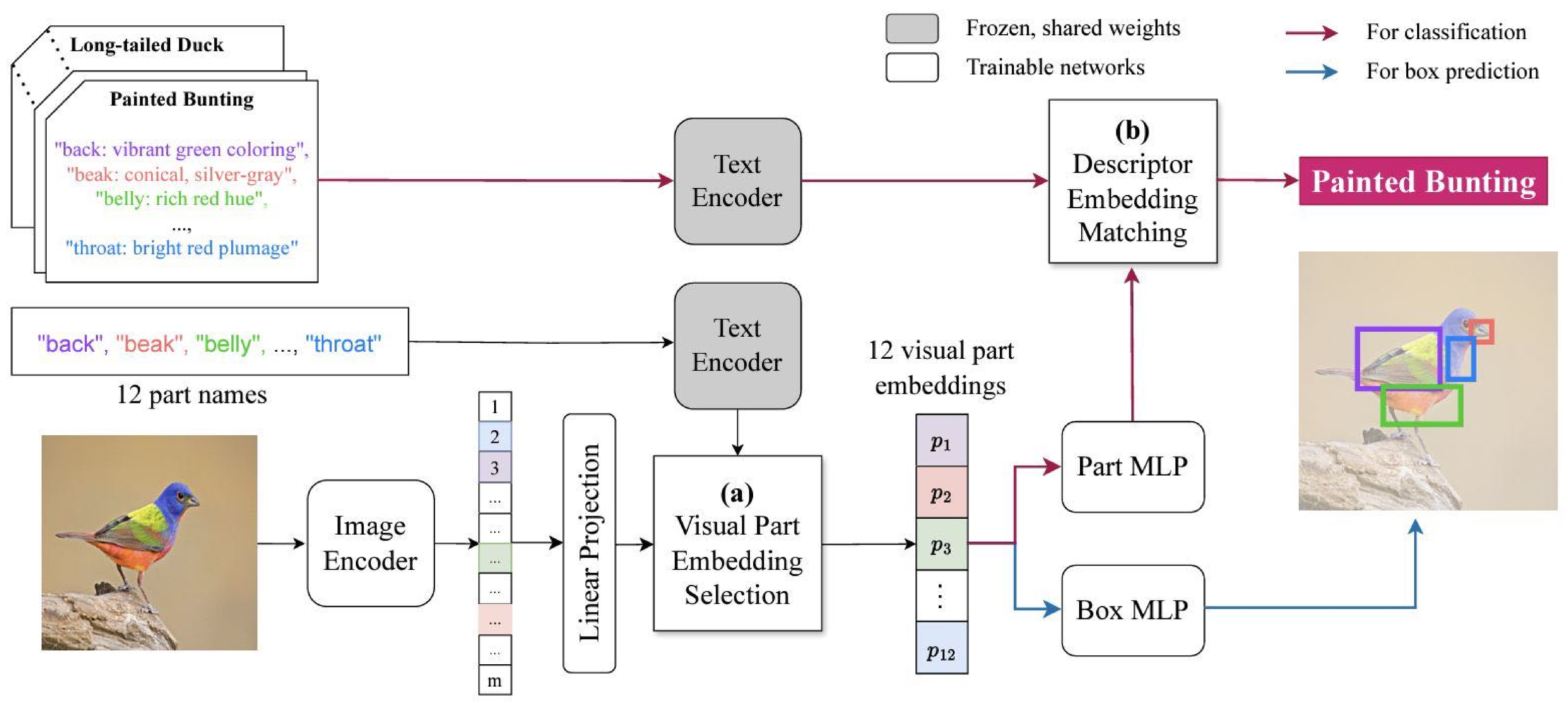
ProtoPNet (2019), ProtoTree (2021), TesNet (2021), Deformable ProtoPNet (2022)



part-based prototypes represent image patches but not editable by humans



3. PEEB Architecture

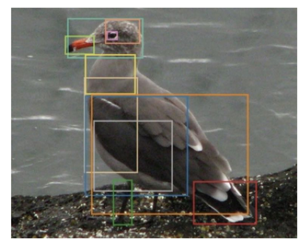


4. How to Train PEEB?

Step 1: Define parts of interest that human experts use for identification

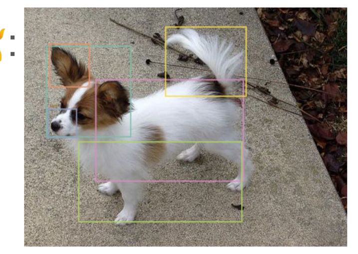
12 parts for birds:

back, beak, belly, breast, crown, forehead, eyes, legs, wings, nape, tail, and throat



6 parts for dogs:

head, body, leg, tail, muzzle, and ear.



Step 2: Prompt GPT-4 for descriptors

- A **bird** has **12 parts**: back, beak, belly, breast, crown, forehead, eyes, legs, wings, nape, tail and throat. Visually describe all parts of **{class name} bird** in a short phrase in bullet points using the format 'part: short phrase'
- A **dog** has **6 parts**: head, body, legs, tail, muzzle and ears. Visually describe all parts of **{class name} dog** in a short phrase in bullet points using the format 'part: short phrase'

```
"cardinal":[
    "back: vibrant red feathers",
    "beak: short, strong, orange",
    "belly: reddish-brown plumage",
    "breast: bright red chest feathers",
    "crown: striking red crest",
    "forehead: vivid red coloration",
    "eyes: small, black, watchful",
    "legs: slender, grey, clawed",
    "wings: red, with black outlines",
    "nape: reddish back of the head",
    "tail: long, red, fan-shaped",
    "throat: rich red plumage"
]
"Toy Poodle":[
    "head: round, small with a soft, gentle expression",
    "ears: long, set high, feathered and hanging close to the head",
    "muzzle: short, square, and deep with a well-defined stop",
    "body: compact but well-proportioned with a level topline",
    "legs: moderate length, straight and with well-feathered fur",
    "tail: docked, carried level with the back and adorned with feathered fur"
]
```

Step 3: Collect data for large-scale pre-training – Bird-11K

Filtering process

- OWL-ViT → bird bounding boxes → removed if the boxes < 100x100 pixel.
- General class names (e.g., Cardinal) are removed while specific ones are kept (e.g., Yellow Cardinal or Northern Cardinal)

Data splits

- GZSL: Excluding test sets only (images)
- ZSL: Excluding all classes (images + descriptors)

Dataset	# of **Images**	# of **Species**
CUB-200-2011 (Wah et al., 2011)	12,000	200
Indian Birds (Vaibhav Rokde, 2023)	37,000	25
NABirds v1 (Van Horn et al., 2015)	48,000	400
Birdsnap v7 (Berg et al., 2014)	49,829	500
iNaturalist 2021-birds (Van Horn et al., 2021)	74,300	1,320
ImageNet-birds (Deng et al., 2009)	76,700	59
BIRDS 525 (Piosenka, 2022)	89,885	525
Macaulay Library at the Cornell Lab of Ornithology	55,283	10,534
Bird-11K (Raw Data)	440,934	11,097
Bird-11K (pre-training set)	294,528	10,811

Step 4: Pre-train PEEB to (1) match image to part descriptors and (2) detect object parts

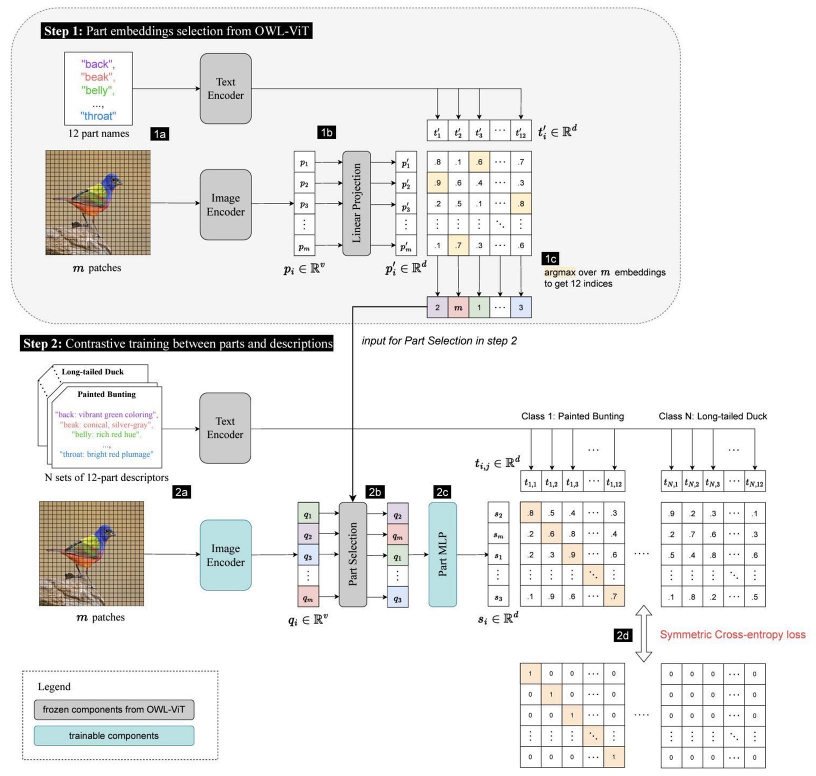
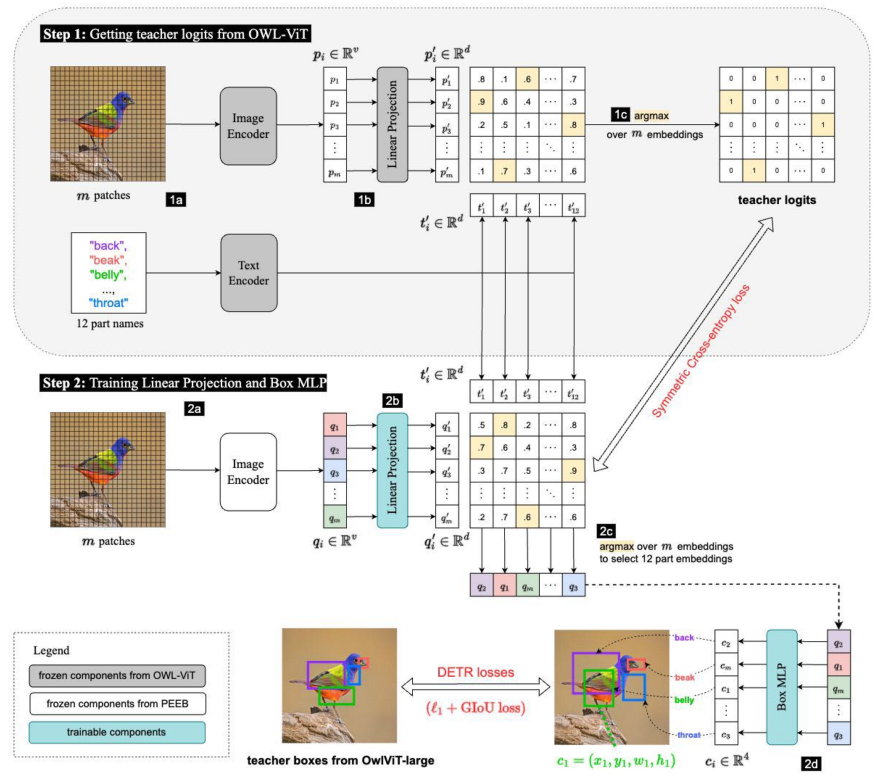


image to part descriptors



object parts detection

6. Conclusion

- PEEB – an explainable and editable classifier that grounds part descriptors to visual bird/dog parts for more fine-grained explanations
- PEEB outperforms CLIP- and text concept-based methods in the zero-shot (ZSL) and generalized zero-shot (GZSL) settings.
- After fine-tuning, PEEB achieves comparable performance to SOTA black-box classifiers.
- PEEB is applicable to other domains (e.g., dogs).

5. Experiments & Results

#1. CLIP-based classifiers depend mostly on class names (not part descriptors)

Table 1: Top-1 test accuracy (%) on CUB-200 when using original, correct (a) vs. randomized, wrong descriptors (b). See Fig. 4 for an example of the descriptors.

	CLIP (2021)	M&V (2023)		PEEB
With class names	✓	✓	✗	✗
(a) Original descriptors	52.02	53.78	5.89	**64.33**
(b) Randomized descriptors	n/a	52.88	0.59	0.88

Table 2: In the **GZSL** setting, PEEB outperforms CLIP and M&V by a large margin, from +8 to +29 pp in top-1 accuracy (see Sec. 5.3). PEEB is also ~10× better than the other two models when class names are replaced by scientific names. As PEEB does not use class names, its accuracy remains unchanged when class names are changed into the scientific ones.

Acc (%)	CUB-200		NABirds-555		iNaturalist-1486	
CLIP (2021)	52.02	(5.95)	39.35	(4.73)	16.36	(2.03)
M&V (2023)	53.78	(7.66)	41.01	(6.27)	17.57	(2.87)
PEEB (**ours**)	**64.33**	(64.33)	**69.03**	(69.03)	**25.74**	(25.74)

#2. Pre-trained PEEB outperforms CLIP-based classifiers and text concept-based classifiers on GZSL setting; and generalizes to traditional ZSL

Table 3: PEEB achieves SOTA CUB-200 accuracy among the **text descriptor-based** classifiers in GZSL. * *1-shot learning.* † *k-means with k = 32.*

Method	Acc (%)	{c}	Textual descriptors
(a) Vision-language models with class names {c} in the prompt			
CLIP (2021)	52.02	✓	Image-level
M&V (2023)	53.78	✓	Image-level
FuDD (2023)	54.30	✓	Image-level
Han et al. (2023b)	56.13	✓	Image-level
(b) Vision-language models with text bottlenecks and no class names {c}			
LaBo (2023)	54.19†	✗	Image-level
Yan et al. (2023)	60.27*	✗	Image-level, attribute-based
PEEB (**ours**)	**64.33**	✗	Part-level
GPT-4V (2023)	69.40	✓	Part-level
(c) Concept-Bottleneck Models with attribute-based, non-textual bottlenecks			
CBM (2020)	62.90	✗	Attribute-based, tabular data
PCBM (2023)	61.00	✗	Attribute-based, tabular data

Table 4: PEEB consistently outperforms other vision-language methods under Harmonic mean and especially in the hard split (SCE) by (+5 to +15) points, highlighting its generalization capability on ZSL.

Methods	CUB Seen	CUB Unseen	CUB Mean	NABirds Seen	NABirds Unseen	NABirds Mean
(a) Data split by Akata et al. (2015)						
CLORE$_{CLIP}$ (2023a)	65.80	39.10	49.05	n/a		
PEEB (**ours**)	**80.78**	**41.74**	**55.04**			
(b) SCS/SCE splits by Elhoseiny et al. (2017)						
	SCS (Easy)	SCE (Hard)	Mean	SCS (Easy)	SCE (Hard)	Mean
S²GA-DET (2018)	42.90	10.90	17.38	39.40	9.70	15.56
GRZSL (2018)	44.08	14.46	21.77	36.36	9.04	14.48
ZEST (2020)	**48.57**	15.26	23.22	38.51	10.23	16.17
CANZSL (2020)	45.80	14.30	21.12	38.10	8.90	14.43
DGRZSL (2021)	45.48	14.29	21.75	37.62	8.91	14.41
DPZSL (2023)	45.40	15.50	23.11	**40.80**	8.20	13.66
PEEB (**ours**)	44.66	**20.31**	**27.92**	28.26	**24.34**	**26.15**

#3. Fine-tuning yields competitive explainable classifiers on bird and dog domains

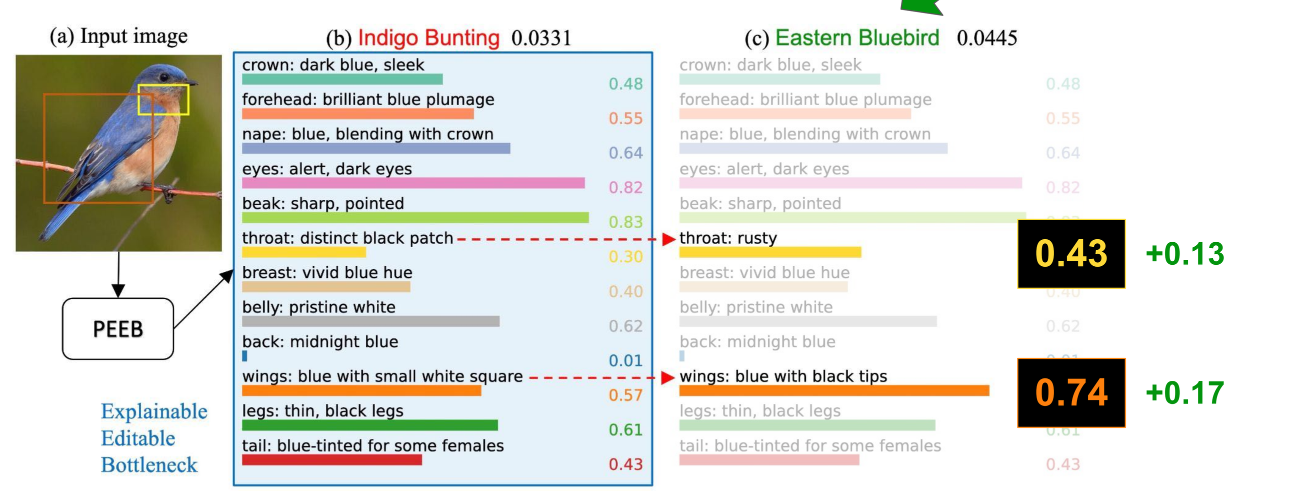
Table 5: PEEB is a state-of-the-art, explainable CUB-200 classifiers in the **supervised** learning.

Methods	Model size	Backbone	Acc (%)
(a) SOTA **black**-box classifiers			
Base (ViT) (2021)	22M	DeiT-S (2021)	84.28
ViT-Net (2022a)	26M	DeiT-S	90.10
(b) Concept-bottleneck classifiers			
CBM (Koh et al., 2020)	11M	ResNet-18	80.10
CPM (Panousis et al., 2023)	155M	ViT-B/16	72.00
CDM (Oikarinen et al., 2023)	155M	ViT-B/16	74.31
LaBo (Yang et al., 2023)	427M	ViT-L/14	81.90
(c) Part-based, explainable classifiers			
ProtoPNet (2019)	22M	DeiT-S	84.04
ProtoTree (2021)	92M	ResNet-50	82.20
TesNet (2021)	79M	DenseNet-121	84.80
Deformable ProtoPNet (2022)	23M	ResNet-50	86.40
ProtoPFormer (2022)	22M	DeiT-S	84.85
PEEB (**ours**)	155M		
~~pre-training +~~ finetuning *only*	155M	OWL-ViT$_{B/32}$	77.80
pre-training + finetuning	155M	OWL-ViT$_{B/32}$	**86.73**
pre-training + finetuning	155M	OWL-ViT$_{B/16}$	**88.80**

Table 6: In the **supervised** learning setting, PEEB is the state-of-the-art explainable, Stanford Dogs-120 classifiers and competitive w.r.t. SOTA black-box models.

Methods	Model size	Backbone	Acc (%)
(a) SOTA **black**-box classifiers			
TransFG (2022a)	86M	ViT-B/16	92.30
ViT-Net (2022b)	86M	DeiT-B	93.60
SR-GNN (2022)	32M	Xception	97.00
(b) Explainable methods			
FCAN (2016)	50M	ResNet-50	84.20
RA-CNN (2017)	144M	VGG-19	87.30
ProtoPNet (2019)	22M	DeiT-S	77.30
Deformable ProtoPNet (2022)	23M	ResNet-50	86.50
PEEB (**ours**)	155M		
~~pre-training +~~ finetuning *only*	155M	OWL-ViT$_{B/32}$	74.17
pre-training ~~+ finetuning~~	155M	OWL-ViT$_{B/32}$	87.37
pre-training + finetuning	155M	OWL-ViT$_{B/16}$	**92.20**

#4. PEEB is editable to add new unseen classes



#5. Qualitative result

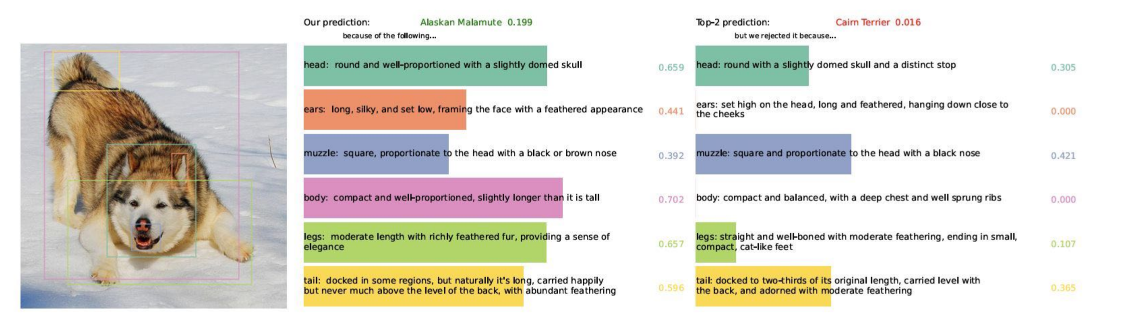


Figure 5: PEEB classifies this Dogs-120 image into `Alaskan Malamute` (softmax: 0.199) due to the matching between the image regions and associated textual part descriptors. In contrast, the explanation shows that the input image is not classified into `Cairn Terrier` mostly because its `ears` and `body` regions do *not* match the text descriptors, i.e., dot products are 0.000 and 0.000, respectively. See Appendix G for more qualitative examples.