南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Generalized Category Discovery

Sagar Vaze* Kai Han† Andrea Vedaldi* Andrew Zisserman*

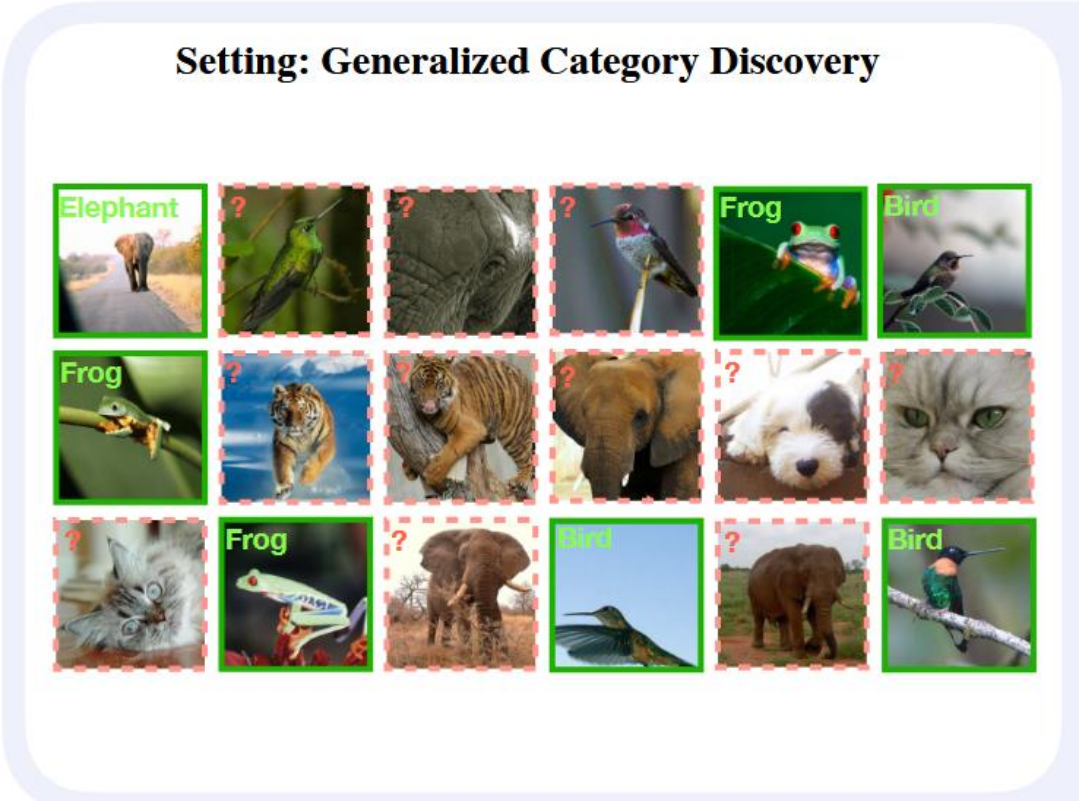
*Visual Geometry Group, Department of Engineering Science, University of Oxford

†The University of Hong Kong

{sagar,vedaldi,az}@robots.ox.ac.uk kaihanx@hku.hk

CVPR 2022

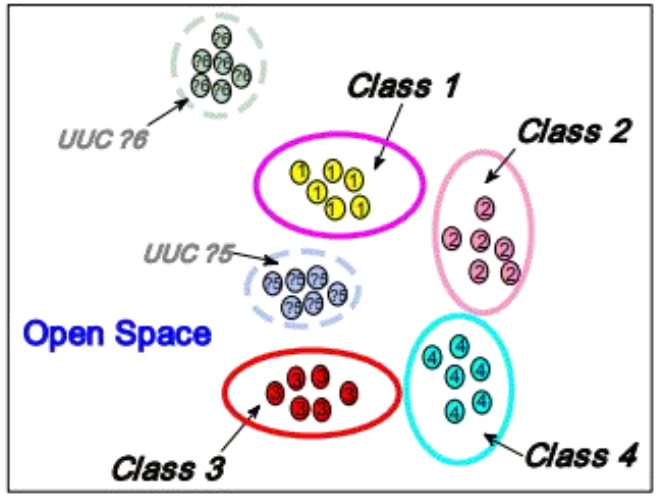
Background



Existing recognition methods are not able to deal with this setting, because they make several restrictive assumptions, such as the unlabelled instances only coming from known – or unknown – classes, and the number of unknown classes being known a-priori. We address the more unconstrained setting, naming it 'Generalized Category Discovery', and challenge all these assumptions.

Background

open-set recognition (OSR)



(c) Open set recognition/classification problem.

novel-category discovery (NCD)

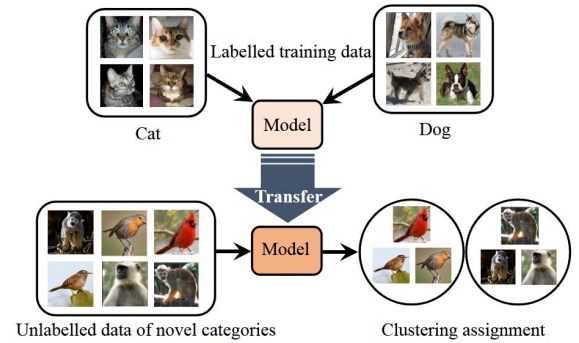
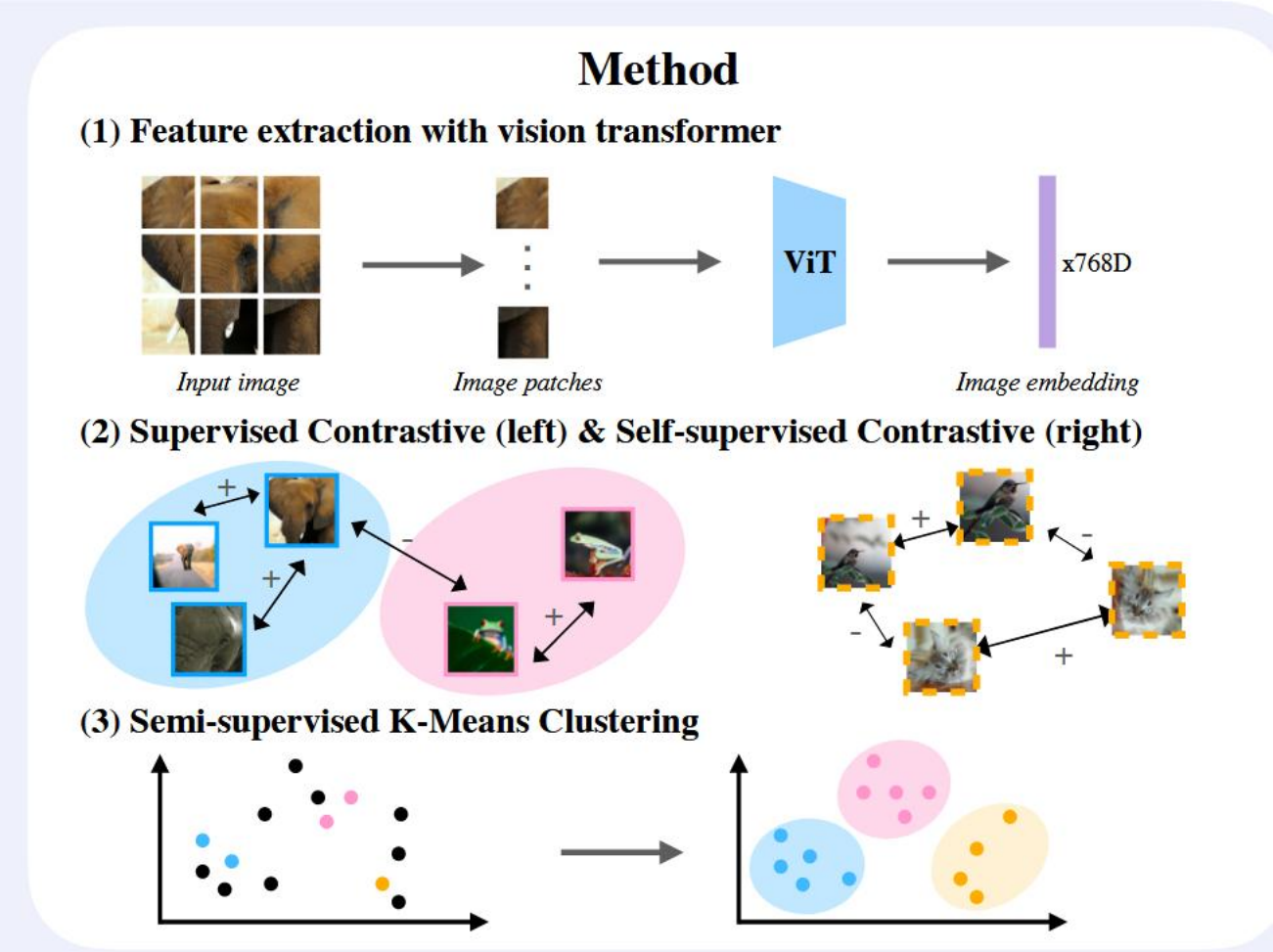


Figure 1. Learning to discover novel visual categories via deep transfer clustering. We first train a model with labelled images (e.g., cat and dog). The model is then applied to images of unlabelled novel categories (e.g., bird and monkey), which transfers the knowledge learned from the labelled images to the unlabelled images. With such transferred knowledge, our model can then simultaneously learn a feature representation and the clustering assignment for the unlabelled images of novel categories.

项目	OSR（开放集识别）	NCD（新类发现）	GNCD（广义新类发现）
是否有标签数据	☑仅包含已知类	❌完全无标签（只包含新类）	☑有标签（已知类）+ 无标签（混合）
核心任务目标	区分“已知类”与“未知类”，未知类拒识	在无标签数据中聚类出新类结构	在混合数据中分类已知类，聚类新类
是否做新类分类	❌不需要	☑聚类发现	☑聚类发现
是否分类已知类	☑是	❌否	☑是
输入测试数据	仅测试阶段含未知类	仅新类无标签数据	混合：已知类+未知类，无标签

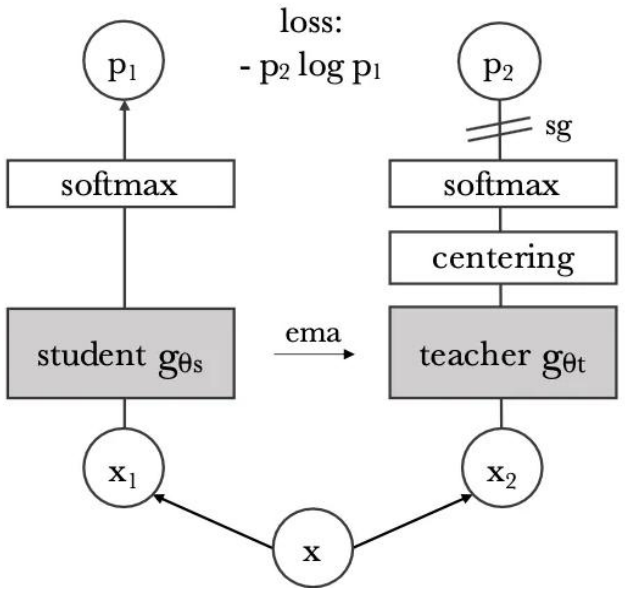
Methods



有标签：$D_{L=}\{(x_i, y_i)\}_{I=1}^{N} \in X\ x\ y_L$　　无标签：$D_{U=}\{(x_i, y_i)\}_{I=1}^{M} \in X\ x\ y_U$

Methods

(1) Representation learning



DINO架构图

① Unsupervised contrastive loss

$$\mathcal{L}_i^u = -\log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_i'/\tau\right)}{\sum_n \mathbb{1}_{[n\neq i]} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_n/\tau\right)}, \tag{1}$$

where $\mathbf{z}_i = \phi(f(\mathbf{x}_i))$ and $\mathbb{1}_{[n\neq i]}$ is an indicator function evaluating to 1 *iff* $n \neq i$, and τ is a temperature value. f is the feature backbone, and ϕ is a multi-layer perceptron (MLP) projection head.

② Supervised contrastive loss

$$\mathcal{L}_i^s = -\frac{1}{|\mathcal{N}(i)|} \sum_{q\in\mathcal{N}(i)} \log \frac{\exp\left(\mathbf{z}_i \cdot \mathbf{z}_q/\tau\right)}{\sum_n \mathbb{1}_{[n\neq i]} \exp\left(\mathbf{z}_i \cdot \mathbf{z}_n/\tau\right)}, \tag{2}$$

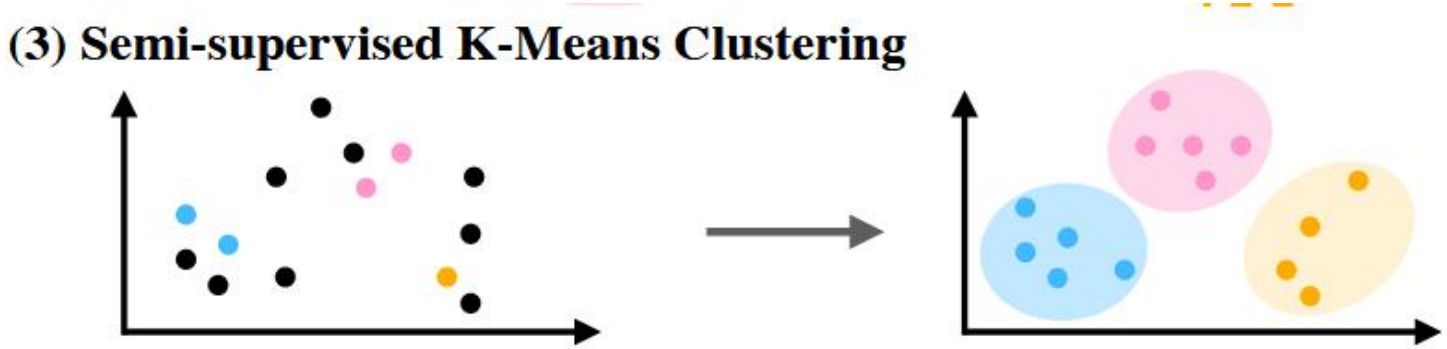
③ Total loss

$$\mathcal{L}^t = (1-\lambda)\sum_{i\in B} \mathcal{L}_i^u + \lambda \sum_{i\in B_{\mathcal{L}}} \mathcal{L}_i^s \tag{3}$$

Methods

(2) Label assignment with semi-supervised k-means

Instead of performing this parametrically as is common in NCD (and risk overfitting to the labelled data) we propose to use a non-parametric method.



Label **kmean**
instances from the same class in DL are always forced to have the same cluster assignment

UnLabel **kmean++**
each instance in DU can be assigned to any cluster based on the distance to different centroids

(3) Estimating the class number in unlabelled data

问题

在 NCD 和无监督聚类设置中，通常假设事先知道数据集中的类别数量，但这在现实世界中是不现实的，因为标签本身是未知的

具体步骤

（1）对数据集进行聚类

- 对整个数据集 D，使用 k-means 聚类算法进行聚类，设定不同的聚类簇数 k
- 每次聚类后，得到每个样本的聚类标签

（2）计算聚类准确率 ACC(k)

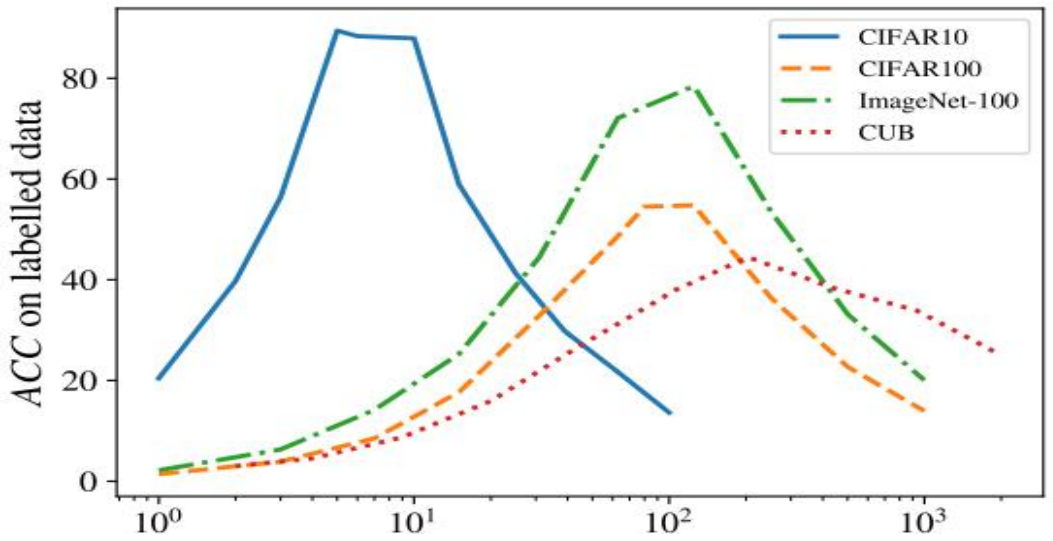
- 只利用带标签的子集 D_L，将其聚类标签与真实标签进行匹配；
- 使用匈牙利算法（Hungarian algorithm）实现聚类标签与真实标签的一对一最优匹配；
- 统计匹配后 D_L 上的分类准确率，作为 ACC(k)

（3）寻找最优聚类k*

- 由于真实类别数未知，通过在区间 $[C_L, C_{max}]$内搜索 k，计算对应的 ACC(k)
- 使用黑盒优化算法（如 Brent 算法）寻找使 ACC(k) 最大的 k*
- 这个 k* 就是对所有类别（已知和未知）的总数的最佳估计

$$k^* = arg\max_k ACC(k)$$

Methods



Total clusters in k-means

Table 1. Datasets used in our experiments. We show the number of classes in the labelled and unlabelled sets ($|\mathcal{Y}_\mathcal{L}|$, $|\mathcal{Y}_\mathcal{U}|$), as well as the number of images ($|\mathcal{D}_\mathcal{L}|$, $|\mathcal{D}_\mathcal{U}|$).

	CIFAR10	CIFAR100	ImageNet-100	CUB	SCars	Herb19
$\lvert\mathcal{Y}_\mathcal{L}\rvert$	5	80	50	100	98	341
$\lvert\mathcal{Y}_\mathcal{U}\rvert$	10	100	100	200	196	683
$\lvert\mathcal{D}_\mathcal{L}\rvert$	12.5k	20k	31.9k	1.5k	2.0k	8.9k
$\lvert\mathcal{D}_\mathcal{U}\rvert$	37.5k	30k	95.3k	4.5k	6.1k	25.4k

Experiments

BaseLines

① **RankStats+**
RankStats trains two classifiers on top of a shared feature representation: the first head is fed instances from the labelled set and is trained with the cross-entropy loss, while the second head sees only instances from unlabelled classes (again, in the NCD setting, the labelled and unlabelled classes are disjoint).
②**UNO+**
Similarly to RankStats, UNO is trained with classification heads for labelled and unlabelled data. The model is then trained in a SwAV-like manner.

Evaluation protocol

$$ACC = \max_{p \in \mathcal{P}(\mathcal{Y}_u)} \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}\{y_i = p(\hat{y}_i)\} \qquad (4)$$

使用 匈牙利算法（Hungarian algorithm） 实现的，它用于在预测簇和真实类之间找出最佳一一对应关系

Experiments

Table 2. Results on generic image recognition datasets.

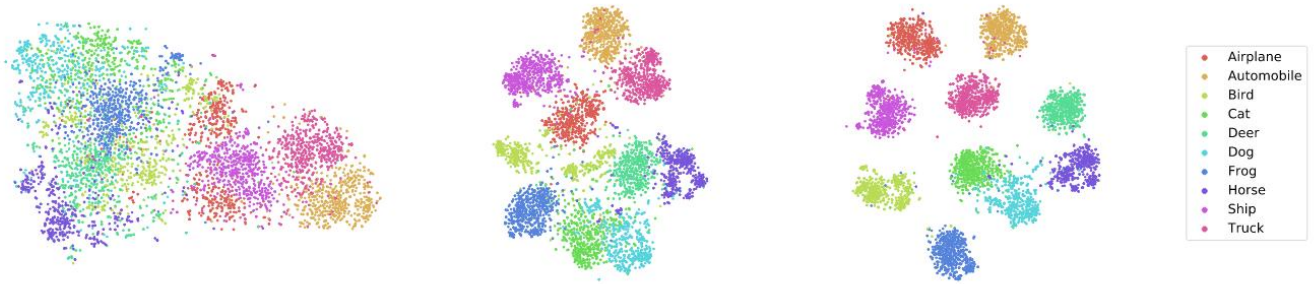
	CIFAR10			CIFAR100			ImageNet-100		
Classes	All	Old	New	All	Old	New	All	Old	New
k-means [30]	83.6	85.7	82.5	52.0	52.2	50.8	72.7	75.5	**71.3**
RankStats+	46.8	19.2	60.5	58.2	77.6	19.3	37.1	61.6	24.8
UNO+	68.6	**98.3**	53.8	69.5	**80.6**	47.2	70.3	**95.0**	57.9
Ours	**91.5**	97.9	**88.2**	**70.8**	77.6	**57.0**	**74.1**	89.8	66.3

Table 3. Results on SSB [45] and Herbarium19 [42].

	CUB			Stanford Cars			Herbarium19		
Classes	All	Old	New	All	Old	New	All	Old	New
k-means [30]	34.3	38.9	32.1	12.8	10.6	13.8	12.9	12.9	12.8
RankStats+	33.3	51.6	24.2	28.3	61.8	12.1	27.9	**55.8**	12.8
UNO+	35.1	49.0	28.1	35.5	**70.5**	18.6	28.3	53.7	14.7
Ours	**51.3**	**56.6**	**48.7**	**39.0**	57.6	**29.9**	**35.4**	51.0	**27.0**

Table 4. Estimation of the number of classes in unlabelled data.

	CIFAR10	CIFAR100	ImageNet-100	CUB	SCars	Herb19
Ground truth	10	100	100	200	196	683
Ours	9	100	109	231	230	520
Error	10%	0%	9%	16%	15%	28%



ResNet50 (DINO) **ViT (DINO)** **ViT (Ours)**

Figure 2. TSNE visualization of instances in CIFAR10 for features generated by a ResNet-50 and ViT model trained with DINO self-supervision on ImageNet, and a ViT model after fine-tuning with our approach.

Table 5. Ablation study on the different components of our approach.

	ViT Backbone	Contrastive Loss	Sup. Contrastive Loss	Semi-Sup k-means	CIFAR100			Herbarium19		
					All	Old	New	All	Old	New
(1)	✗	✗	✗	✗	34.0	34.8	32.4	12.1	12.5	11.9
(2)	✓	✗	✗	✗	52.0	52.2	50.8	12.9	12.9	12.8
(3)	✓	✓	✗	✗	54.6	54.1	53.7	14.3	15.1	13.9
(4)	✓	✗	✓	✗	60.5	72.2	35.0	17.8	22.7	15.4
(5)	✓	✓	✓	✗	71.1	**78.3**	56.6	28.7	32.1	26.9
(6)	✓	✓	✓	✓	**73.0**	76.2	**66.5**	**35.4**	**51.0**	**27.0**

Semantic Shift Benchmark (SSB, including CUB and Stanford Cars)

Experiments

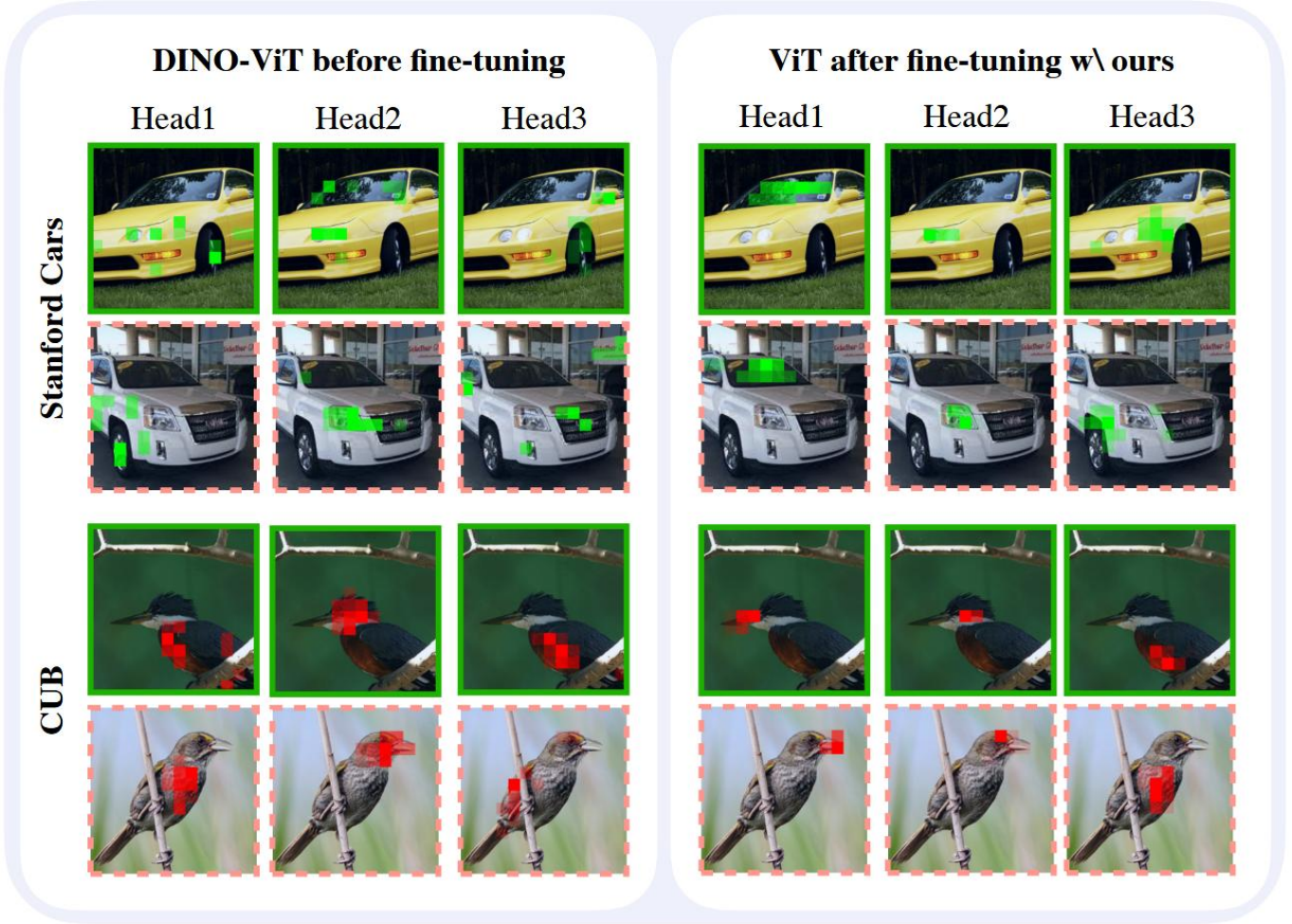


Figure 3. Attention visualizations for the DINO-ViT model before (left) and after (right) fine-tuning with our approach. For Stanford Cars and CUB, we show an image from the 'Old' (first row for each dataset) and 'New' classes (second row for each dataset). Our model learns to specialize attention heads (shown as columns) to different semantically meaningful parts, which can transfer between the labelled and unlabelled categories. The model's heads learn 'Windshield', 'Headlight' and 'Wheelhouse' for the cars, and 'Beak', 'Head' and 'Belly' for the birds. For both models, we select heads with as focused attention as possible. Recommended viewing in color with zoom.