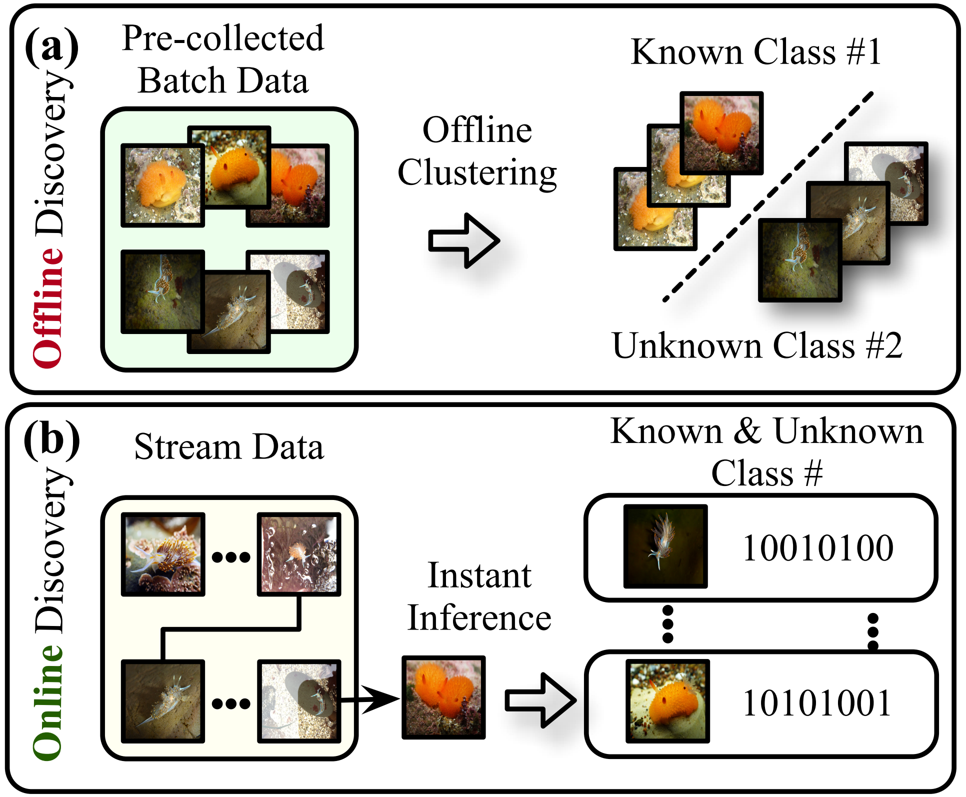
Prototypical Hash Encoding for On-the-Fly Fine-Grained Category Discovery

Haiyang Zheng [1], Nan Pu [1], Wenjing Li [2], Nicu Sebe [1], and Zhun Zhong [2]
[1] University of Trento [2] Hefei University of Technology

Problem: On-the-Fly Category Discovery (OCD)



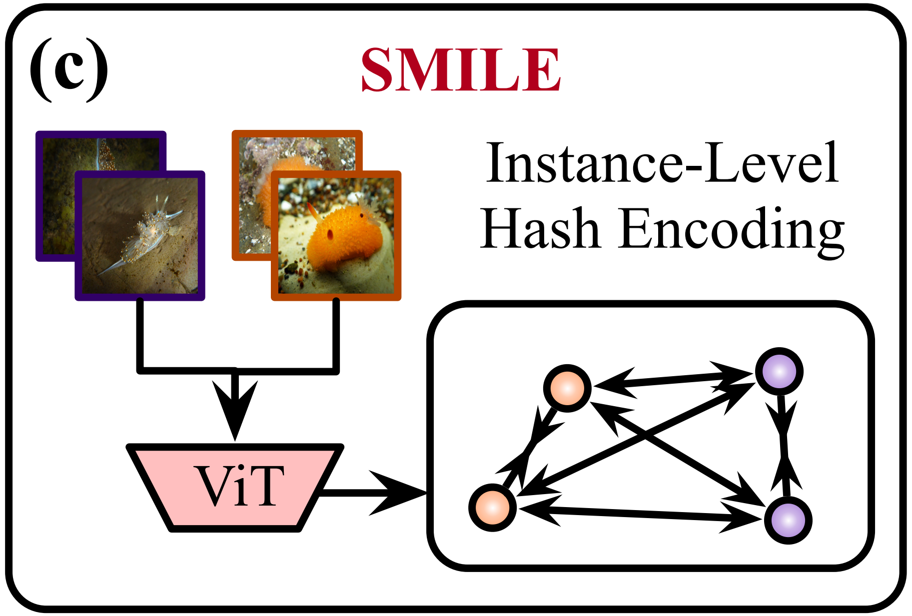
On-the-Fly category discovery (OCD) aims to online discover the newly-coming stream data that belong to both known and unknown classes, by leveraging only known category knowledge contained in labelled data.

Challenges:

Only labelled data is available during training.

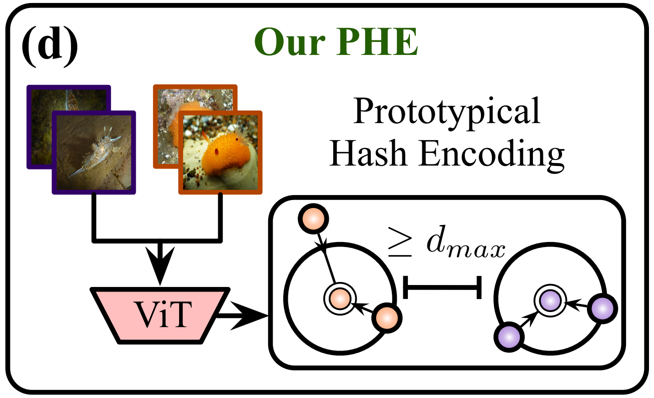
Stream data for instant inference.

Previous works



SMILE directly maps image features into low-dimensional hash space with an instance-level contrastive objective and regard that one hash code uniquely represents a category. Given this, although SMILE can derive category descriptors, it suffers from a significant issue of "high sensitivity" for learned hash-form category descriptors and thus produces a significantly inaccurate number of categories as well as unsatisfied performance.

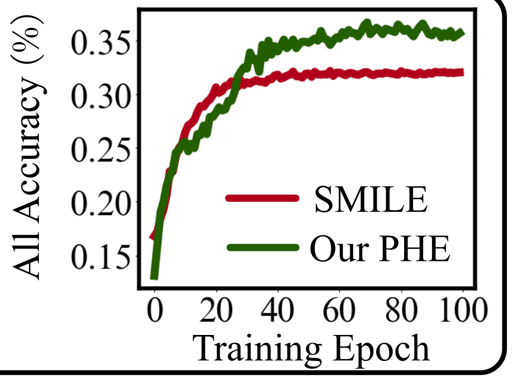
Motivation



(e) Ground Truth # Class = 200

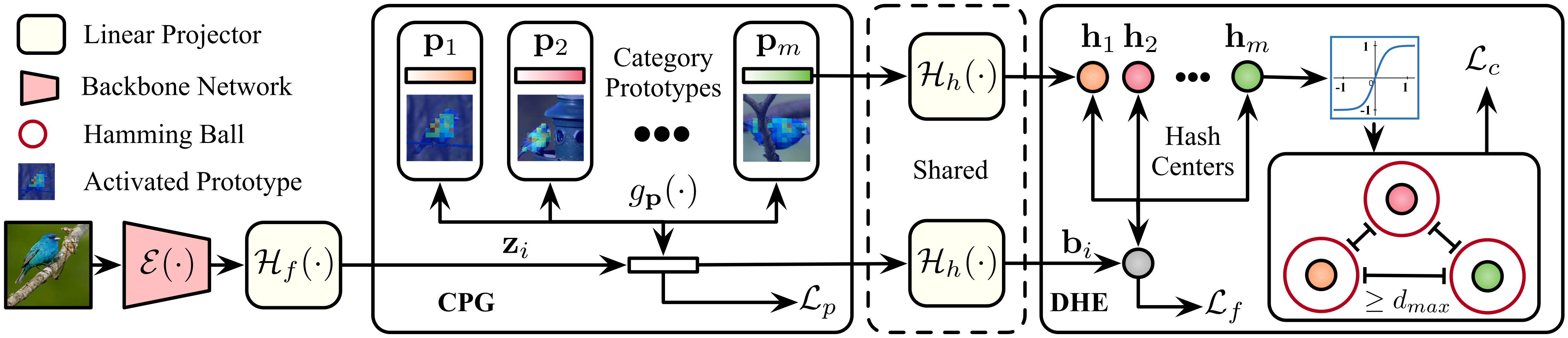
Methods	Code Length	All Acc (%)	Estimated # Class
SMILE	12	32.2	580
Ours	12	36.4	339
SMILE	64	22.6	2910
Ours	64	38.1	493

High Sensitivity



Current methods that map features directly into a low-dimensional hash space not only inevitably **damages the ability to distinguish between classes** but also introduce a **"high sensitivity" issue**, especially for fine-grained classes, leading to inferior performance.

Method



To achieve accurate and online category discovery, we design a **Prototypical Hash Encoding (PHE) framework**, which mainly consists of a **Category-aware Prototype Generation (CPG)** module and a **Discriminative Hash Encoding (DHE)** module. CPG aims at modeling diverse intra-category information and generating category-specific prototypes for representing fine-grained categories. DHE leverages generated prototypical hash centers to further facilitate discriminative hash code generation.

Category-aware Prototype Generation (CPG)

The CPG module employs **Prototype-based Interpretable Models** to generate multiple category prototypes for each fine-grained category, effectively modeling the diverse intra-category information with following loss function.

$$\mathcal{L}_p = \frac{1}{|B|} \sum_{i \in B} \ell(\boldsymbol{y}_i, FC(\mathcal{B}(\theta) \cdot \mathbf{s}_i))$$

Discriminative Hash Encoding (DHE)

The DHE module focuses on **hash-based category encoding**. Image features and category prototypes are mapped into hash codes and hash centers, respectively, through a shared projection layer. We use the following loss to optimize the hash features of the images to be closer to their corresponding hash centers.

$$\mathcal{L}_f = \frac{1}{|B|}\sum_{i \in B} \ell(\boldsymbol{y}_i, sim(\mathbf{b}_i, \mathbf{h}))$$

Discriminative Hash Encoding (DHE)

We design a **center separation loss** to ensure that the Hamming distance between any two hash centers is at least d, thereby guaranteeing inter-class separability. The maximum separation threshold d_{max} is derived from the **Gilbert-Varshamov bound** in coding theory. Additionally, a quantization loss is used. The optimization loss $\mathcal{L}_c = \mathcal{L}_{sep} + \mathcal{L}_q$.

$$\mathcal{L}_{sep} = \sum_i \sum_j \max(0, d - ||\hat{\mathbf{h}}_i - \hat{\mathbf{h}}_j||_H) \quad , \quad \mathcal{L}_q = \sum_i (1 - |\hat{\mathbf{h}}_i|)$$

Training and Inference

Model Training. During the model training process, the total loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_p + \alpha * \mathcal{L}_c + \beta * \mathcal{L}_f,$$

where α and β control the importance of center optimization and hash encoding, respectively.

Hamming Ball Based Model Inference. During on-the-fly testing, given an input image x_i in the query set D_Q, we use $\hat{\mathbf{b}}_i = sign(\mathcal{H}_h(\mathcal{H}_f(\mathcal{E}(\mathbf{x}_i))))$ as its category descriptor. Due to the introduction of the center separation loss, the Hamming distance between any two hash centers is not less than d_{max}. We consider a Hamming ball centered on the hash centers with a radius of $\max(\lfloor \frac{d_{max}}{2} \rfloor, 1)$ to represent a category. Specifically, during inference, if the Hamming distance between $\hat{\mathbf{b}}_i$ and any existing hash center is less than or equal to $\max(\lfloor \frac{d_{max}}{2} \rfloor, 1)$, we classify the image as belonging to the corresponding category of that hash center. Otherwise, the image is used to establish a new hash center and category.

Experiment

Achieve a new state-of-the-art performance on eight fine-grained datasets.

Method	CUB			Stanford Cars			Oxford Pets			Food101			Average		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
SLC	31.3	48.5	22.7	24.0	45.8	13.6	35.5	41.3	33.1	20.9	48.6	6.8	27.9	46.1	19.1
RankStat	27.6	46.2	18.3	18.6	36.9	9.7	33.2	42.3	28.4	22.3	50.7	7.8	25.4	44.0	16.1
WTA	26.5	45.0	17.3	20.0	38.8	10.6	35.2	46.3	29.3	18.2	40.5	6.1	25.0	42.7	15.8
SMILE	32.2	50.9	22.9	26.2	46.7	16.3	41.2	42.1	40.7	24.0	54.6	8.4	30.9	48.6	22.1
PHE (Ours)	**36.4**	**55.8**	**27.0**	**31.3**	**61.9**	**16.8**	**48.3**	**53.8**	**45.4**	**29.1**	**64.7**	**11.1**	**36.3**	**59.1**	**25.1**
Method	**Fungi**			**Arachnida**			**Animalia**			**Mollusca**			**Average**		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
SLC	27.7	60.0	13.4	25.4	44.6	11.4	32.4	**61.9**	19.3	31.1	59.8	15.0	29.2	56.6	14.8
RankStat	23.8	50.5	12.0	26.6	51.0	10.0	31.4	54.9	21.6	29.3	55.2	15.5	27.8	52.9	14.8
WTA	27.5	65.6	12.0	28.1	55.5	10.9	33.4	59.8	22.4	30.3	55.4	17.0	29.8	59.1	15.6
SMILE	29.3	64.6	13.6	29.9	57.9	12.2	35.9	49.4	30.3	33.3	44.5	**27.2**	32.1	54.1	20.8
PHE (Ours)	**31.4**	**67.9**	**15.2**	**37.0**	**75.7**	**12.6**	**40.3**	55.7	**31.8**	**39.9**	**65.0**	26.5	**37.2**	**66.1**	**21.5**

Ablation Study & Encoding Length Evaluation

Ablation study.

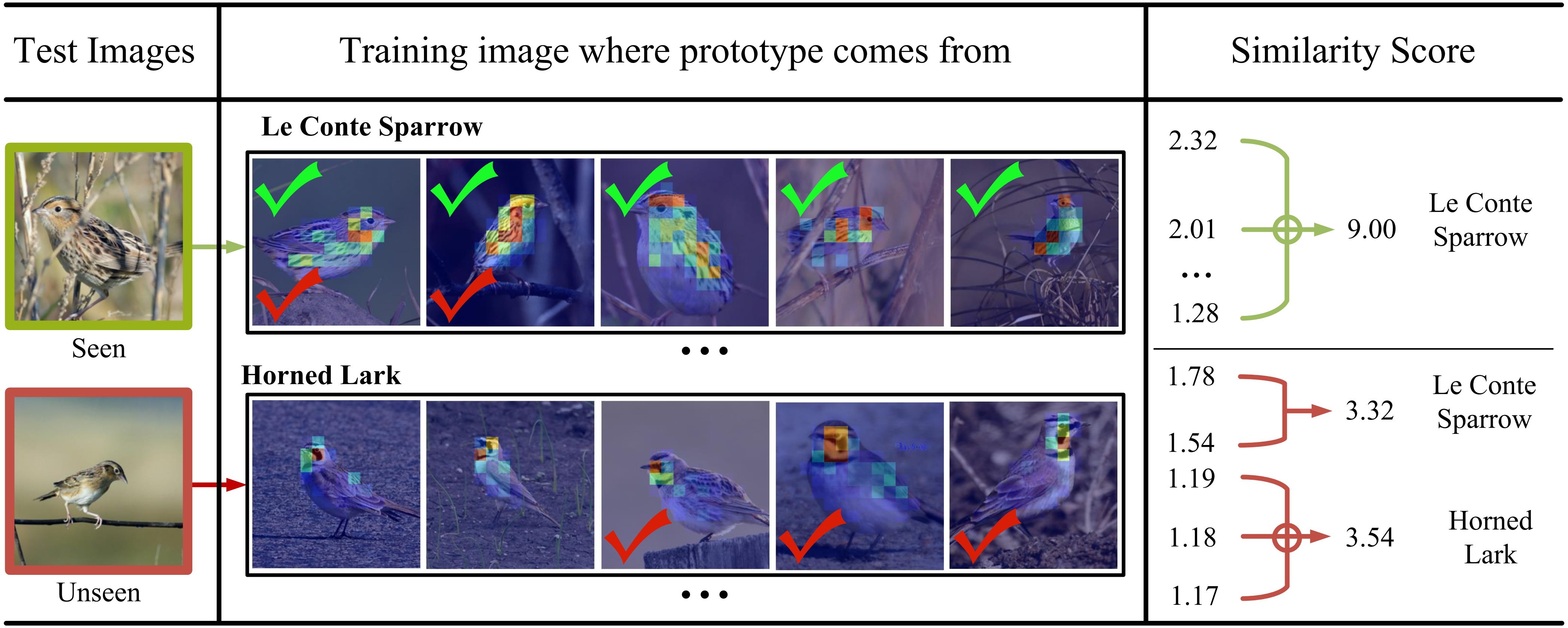
$\mathcal{L}_p$	$\mathcal{L}_c$	$\mathcal{L}_f$	CUB			SCars		
			All	Old	New	All	Old	New
	✓	✓	34.9	53.0	25.8	28.9	58.4	14.6
✓		✓	32.0	43.4	26.4	24.1	40.2	16.3
✓	✓		34.1	54.3	24.0	26.0	52.6	13.1
✓	✓	✓	**36.4**	**55.8**	**27.0**	**31.3**	**61.9**	**16.8**

Performance evaluation with different encoding length.

L	Methods	CUB#200			Estimated	SCars#196			Estimated
		All	Old	New	#Class	All	Old	New	#Class
16bit	SMILE	31.9	52.7	21.5	924	27.5	52.5	15.4	896
	Ours	**37.6**	**57.4**	**27.6**	**318**	**31.8**	**65.4**	**15.6**	**709**
32bit	SMILE	27.3	52.0	14.97	2146	21.9	46.8	9.9	2953
	Ours	**38.5**	**59.9**	**27.8**	**474**	**31.5**	**64.0**	**15.8**	**762**
64bit	SMILE	22.6	45.3	11.2	2910	16.5	38.2	6.1	4788
	Ours	**38.1**	**60.1**	**27.2**	**493**	**32.1**	**66.9**	**15.3**	**917**

Visualization – Case Study

Why is a Grasshopper Sparrow classified as a new category?



Conclusion

In this paper, we introduce a **Prototypical Hash Encoding (PHE)** framework for fine-grained On-thefly Category Discovery. Addressing the limitations of existing methods, which struggle with the high sensitivity of hash-form category descriptors and suboptimal feature representation, our approach incorporates a prototype-based classification model. This model facilitates robust representation learning by **developing multiple prototypes for each fine-grained category**. We then map these category prototypes to corresponding hash centers, optimizing image hash features to align closely with these centers, thereby achieving **intra-class compactness**. Additionally, we enhance **inter-class separation** by maximizing the distance between hash centers, **guided by the Gilbert-Varshamov bound**. Experiments on eight fine-grained datasets demonstrate that our method outperforms previous methods by a large margin. Moreover, a visualization study is provided to understand the underlying mechanism of our method.