THU-AM-276

Open-set Fine-grained Retrieval via Prompting Vision-Language Evaluator

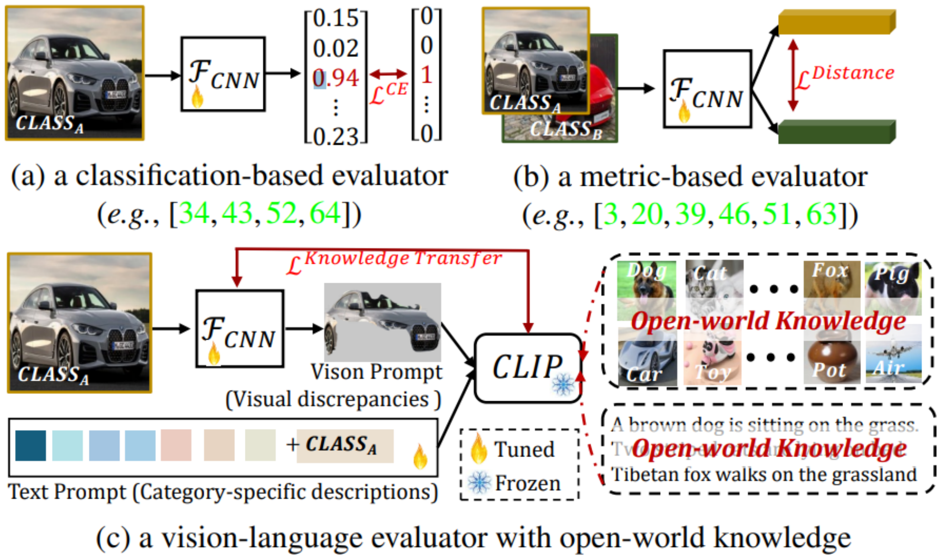
Shijie Wang[1], Jianlong Chang[2], Haojie Li[1,3*], Zhihui Wang[1], Wanli Ouyang[4], Qi Tian[2]
[1]International School of Information Science & Engineering, Dalian University of Technology, China
[2] Huawei Cloud & AI, China
[3]College of Computer and Engineering, Shandong University of Science and Technology, China
[4] Sense Time Computer Vision Research Group, The University of Sydney, Australia

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

DALIAN UNIVERSITY OF TECHNOLOGY 1949 大连理工大学

HUAWEI

Motivation



Problem

Current works focus on close-set visual concepts, where all the subcategories are pre-defined, and make it hard to capture discriminative knowledge from unknown subcategories, consequently failing to handle unknown subcategories in open-world scenarios.

Method

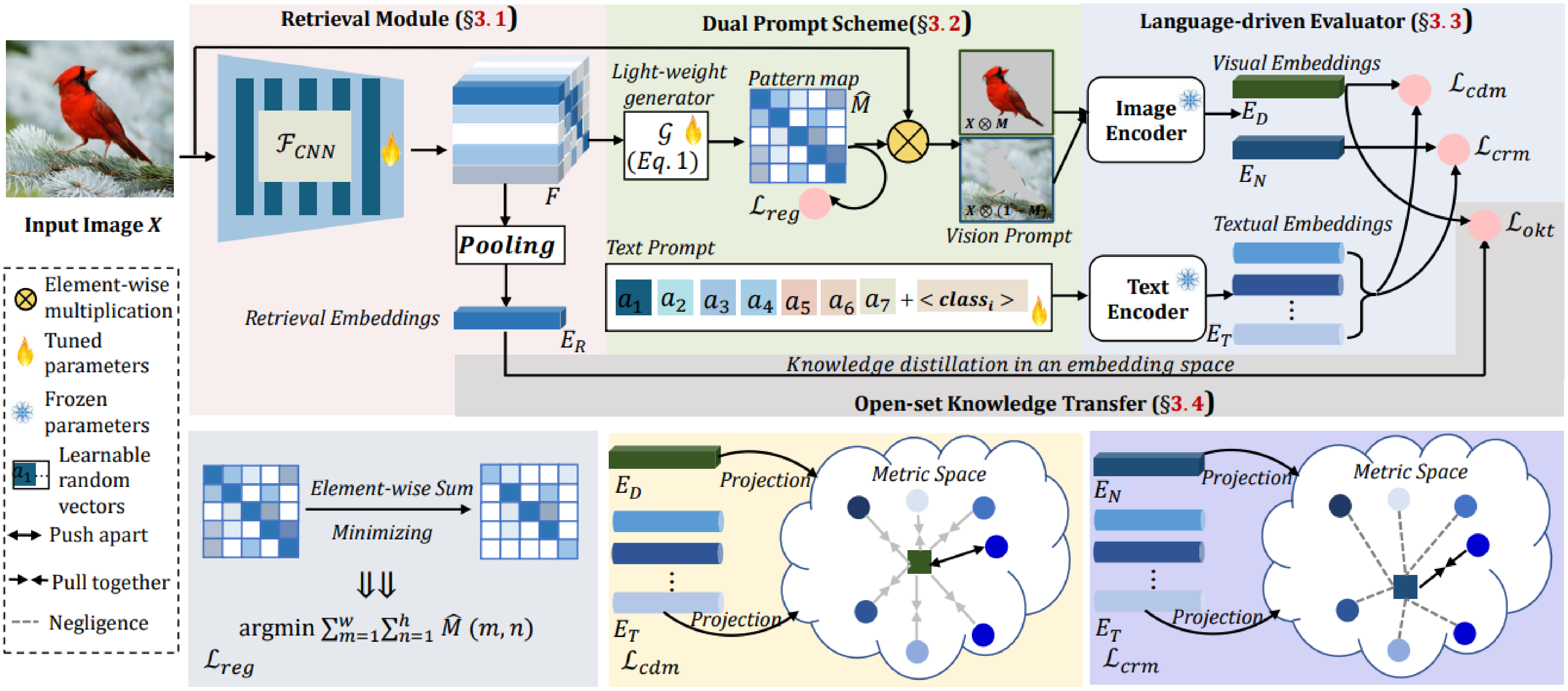
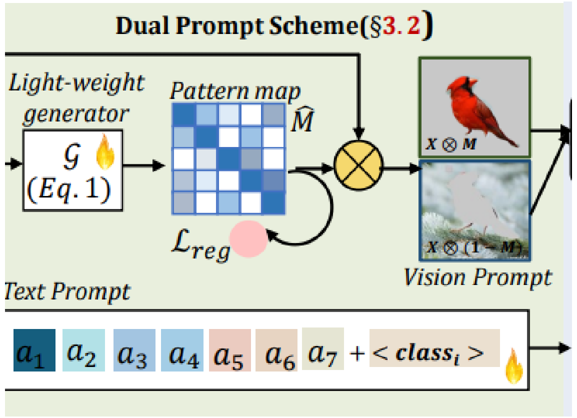


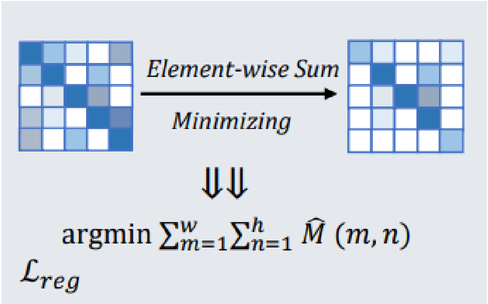
Figure 2. Detailed illustration of **prompting vision-language evaluator**. See §3 for more details.

Experiments

Method	Arch	CUB-200-2011				Stanford Cars 196				FGVC Aircraft			
		1	2	4	8	1	2	4	8	1	2	4	8
SCDA $_{\text{TIP}_{17}}$ [52]	R50	57.3	70.2	81.0	88.4	48.3	60.2	71.8	81.8	56.5	67.7	77.6	85.7
CRL $_{\text{IJCAI}_{18}}$ [64]	R50	62.5	74.2	82.9	89.7	57.8	69.1	78.6	86.6	61.1	71.6	80.9	88.2
CEP $_{\text{ECCV}_{20}}$ [3]	R50	69.2	79.2	86.9	91.6	89.3	93.9	96.6	98.1	-	-	-	-
HDCL $_{\text{IJON}_{21}}$ [59]	R50	69.5	79.6	86.8	92.4	84.4	90.1	94.1	96.5	71.1	81.0	88.3	93.3
DGCRL $_{\text{AAAI}_{19}}$ [65]	R50	67.9	79.1	86.2	91.8	75.9	83.9	89.7	94.0	70.1	79.6	88.0	93.0
DCML $_{\text{CVPR}_{21}}$ [62]	R50	68.4	77.9	86.1	91.7	85.2	91.8	96.0	98.0	-	-	-	-
DAS $_{\text{ECCV}_{22}}$ [27]	R50	69.2	79.3	87.1	92.6	87.8	93.2	96.0	97.9	-	-	-	-
IBC $_{\text{ICML}_{21}}$ [41]	R50	70.3	80.3	87.6	92.7	88.1	93.3	96.2	98.2	-	-	-	-
NIA $_{\text{CVPR}_{22}}$ [40]	R50	70.5	80.6	-	-	89.1	93.4	-	-	-	-	-	-
Proxy $_{\text{CVPR}_{21}}$ [17]	BN	71.1	80.4	87.4	92.5	88.3	93.1	95.7	97.5	-	-	-	-
HIST $_{\text{CVPR}_{22}}$ [25]	R50	71.4	81.1	88.1	-	89.6	93.9	96.4	-	-	-	-	-
ETLR $_{\text{CVPR}_{21}}$ [18]	BN	72.1	81.3	87.6	-	89.6	94.0	96.5	-	-	-	-	-
PNCA++ $_{\text{ECCV}_{20}}$ [46]	R50	72.2	82.0	89.2	93.5	90.1	94.5	97.0	98.4	-	-	-	-
Our PLEor	R50	**74.8**	**84.5**	**91.3**	**94.9**	**94.4**	**96.9**	**98.3**	**98.9**	**86.3**	**91.7**	**95.1**	**96.7**

Dual Prompt Scheme





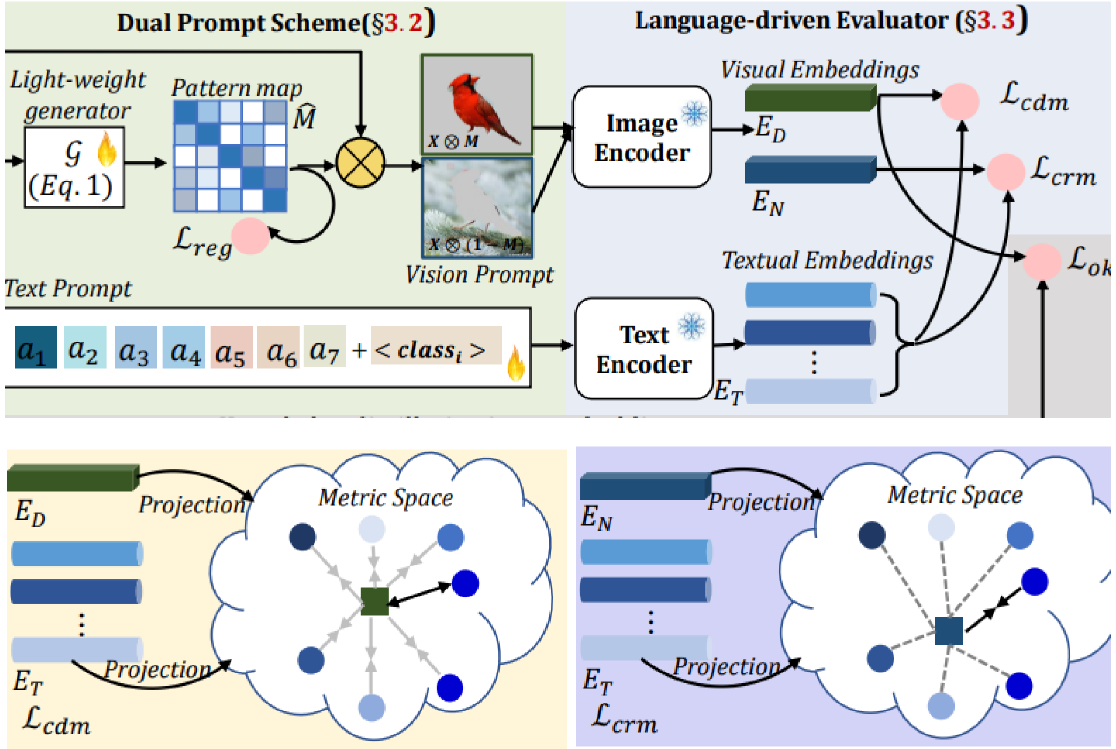
- **Vision Prompt:** To obtain the category-specific discrepancies, the vision prompt aims to project the semantic features into a new space where the location, scale and intensity of these discrepancies are specified.

$$\hat{\mathbf{M}} = \sigma(\mathcal{G}(\mathbf{F})), \quad \mathbf{V_D} = \mathbf{X} \otimes \mathbf{M}, \quad \mathbf{V_N} = \mathbf{X} \otimes (1 - \mathbf{M}),$$

- **Text Prompt:** A text prompt is designed to generate appropriate text descriptions automatically via keeping semantically coherent with the category-specific vision prompt.

$$\mathcal{P}_{\textbf{class}} = (a_1, a_2, \cdots, a_i, \cdots, a_k, < class >),$$

Vision-Language Evaluator



- **Evaluator:** the contrastive objective of vision-language evaluator encourages the pre-trained CLIP model to locate the category-specific descriptions in vision prompt and generate the category-specific semantics into text prompt.

$$\mathcal{L}_{cdm} = -\sum_{i=1}^{N} y_i \cdot log \frac{exp(cos < \mathbf{E_D}, \mathbf{E_T^i} > /\tau)}{\sum_{i=1}^{N} exp(cos < \mathbf{E_D}, \mathbf{E_T^i} > /\tau)},$$

$$\mathcal{L}_{crm} = -\sum_{i=1}^{N} y_i \cdot log(1 - \frac{exp(cos < \mathbf{E_N}, \mathbf{E_T^i} > /\tau)}{\sum_{i=1}^{N} exp(cos < \mathbf{E_N}, \mathbf{E_T^i} > /\tau)}). \quad (7)$$

Ablation Experiments

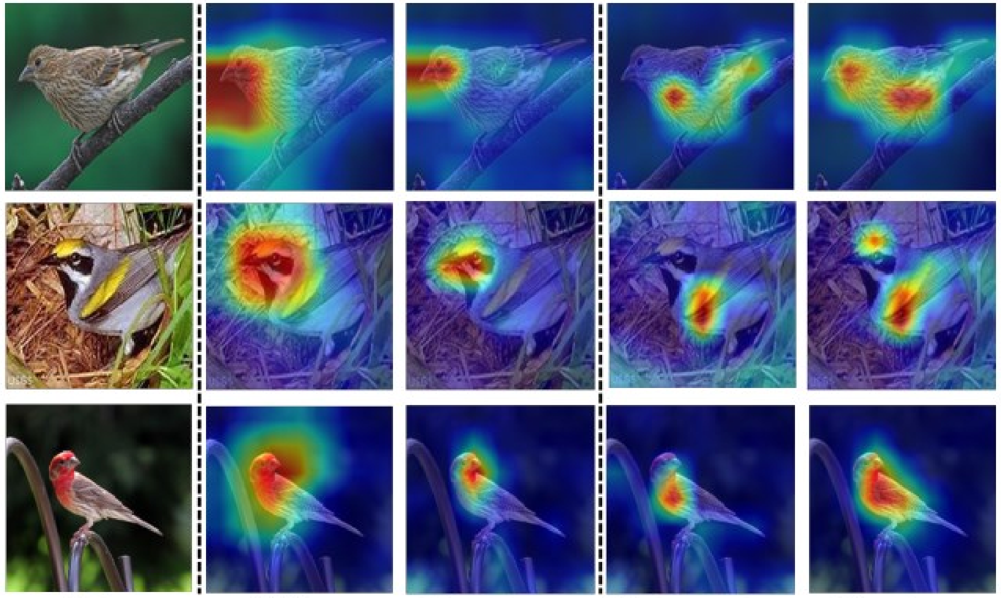
Table 1. Comparison of performance and efficiency on CUB-200-2011 using different combinations of constraints. The first row indicates that we use the tranditional classification-based classifier (*i.e.*, ResNet-50) as supervision, to replace the proposed PLEor for comparison. ”Time” is the time of extracted retrieval embeddings.

$\mathcal{L}_{cdm}$	$\mathcal{L}_{crm}$	$\mathcal{L}_{reg}$	$\mathcal{L}_{okt}$	Recall@1	Time
				66.3%	21.1ms
✓				72.1%	42.3ms
✓	✓			74.4%	42.3ms
✓	✓	✓		75.1%	42.3ms
✓	✓	✓	✓	**74.8%**	**21.1ms**

Table 2. Evaluation results of retrieval performance on CUB-200-2011 dataset with/without the prompt learning. Hand-craft prompt denotes that we use the handcrafted prompt template (”a photo of a [·].”) in text prompt.

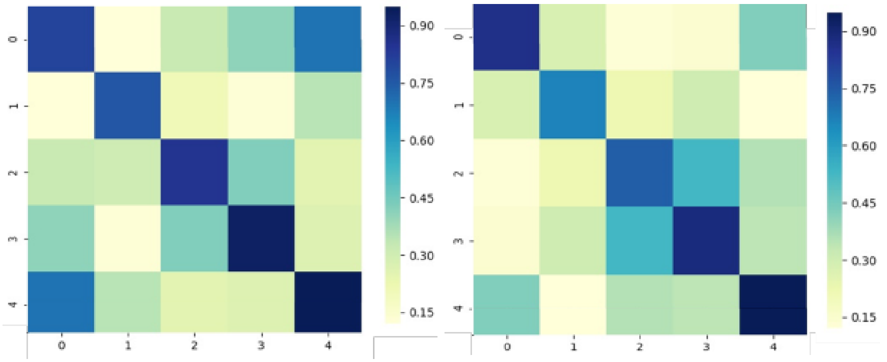
Prompt	Recall@1
CLIP + Hand-craft prompt	71.5%
CLIP + Text Prompt	73.3%$_{+1.8}$
CLIP + Vision&Hand-craft Prompt	72.4%$_{+0.9}$
CLIP + Vision&Text Prompt	**74.8%**$_{+3.3}$

Visualization Experiments



(a) Input (b) $\mathcal{L}_{cls}$ (c) $+\mathcal{L}_{reg}$ (d) $\mathcal{L}_{cdm}+\mathcal{L}_{reg}$ (e) $+\mathcal{L}_{crm}$

Figure 4. Visualization of vision prompt based on classification-based evaluator (b)(c) and our vision-language evaluator (d)(e), respectively. $+\mathcal{L}$ means that we successively add this constraint, *i.e.*, $+\mathcal{L}_{reg}= \mathcal{L}_{cls}+\mathcal{L}_{reg}$, $+\mathcal{L}_{crm}= \mathcal{L}_{cdm}+\mathcal{L}_{reg}+\mathcal{L}_{crm}$.



(a) CUB-200-2011 Train Set (b) CUB-200-2011 Test Set

Figure 5. The nearest description for text prompt learned by PLEor, with their similarity shown in grids.

Conclusion

- A prompting vision-language evaluator, i.e., PLEor, is proposed. It can distill the knowledge with open-world visual concepts from CLIP model to alleviate the problems behind open-set scenarios. To our best knowledge, we are the first to regard CLIP model as an evaluator specifically for OSFR task.

- PLEor provides timely insights into the adaptation of pre-trained CLIP model adopting prompt learning, and crucially, demonstrates the effectiveness of a simple modification for inputs of CLIP model in OSFR.

- PLEor achieves new state-of-the-art results compared with classification-based and metric-based evaluators, which is significant gains of 8.0% average retrieval accuracy on three widely-used OSFR datasets.