# Exploration of visual prompt in Grounded pre-trained open-set detection

Qibo Chen, Weizhong Jin, Shuchang Li, Mengdi Liu, Li Yu, Jian Jiang, Xiaozheng Wang
China Mobile(Zhejiang) Research & Innovation Institute
Emails: chenqibo@tju.edu.cn

## INTRODUCTION

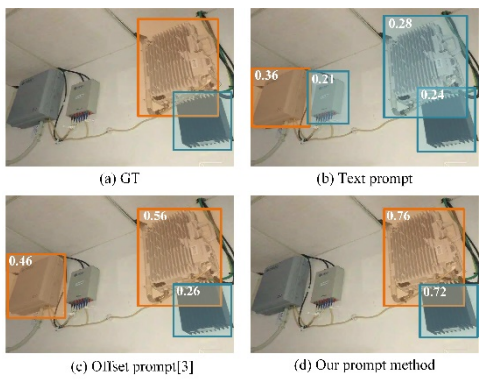
### Limitations of pre-trained open-set detector inference

Most models achieve open-set detection through similarity between text and image regions. But the text prompt suffers from descriptive difficulties, linguistic ambiguity, and adjustment uncertainty in real-world inference, as show in Fig.1.

### Prompt tuning

A direct idea is to optimize the text prompt vector in downstream tasks through GT. However:

- Established methods[2,3] rely on text initialization, limiting the visual prompt representation capabilities.
- The training process does not consider the negative sample problem on the prompt side, and prompts trained from different tasks will lead to a significant drop in precision on when combining inference, as show in Fig.1 (c).



(a) GT (b) Text prompt (c) OffSet prompt[3] (d) Our prompt method

**Figure 1:** Visualization of different prompt combined inference.

### Contribution

- a statistical-based visual prompt construction method that is not dependent on text descriptions initialization.
- a similarity dictionary strategy to make the learned image prompt more discriminative and significantly improve the stability of combined inference.
- Our visual prompt method outperforms ot-her prompt learning methods on 13 public datasets and in combined inference.

## METHOD

### Visual Prompt Construction

visual prompt $\boldsymbol{F_v} = \{\boldsymbol{E_i} \dots \boldsymbol{E_N}\}, \boldsymbol{E} \in \boldsymbol{R^{1\times C}}$ initialized by gaussian distributions, which from pre-training data statistics. Applying stochastic similarity layer to establish similarity, The process defined as

$$\boldsymbol{E_i^{\sim}} = \boldsymbol{aE_i} + \sqrt{\boldsymbol{1-a^2}}\boldsymbol{E_j}\,,\ \boldsymbol{u_1} = \frac{\sum \boldsymbol{E_i}}{\boldsymbol{N}}\,,\ \boldsymbol{u_2} = \frac{\sum \boldsymbol{E_i^{\sim}}}{\boldsymbol{N}},\ \boldsymbol{F_v} \mathrel{-}= (\boldsymbol{u_2} - \boldsymbol{u_1})$$

$\boldsymbol{a} \in [\boldsymbol{0}, \boldsymbol{1}]$ is a constant, controlling similarity between $\boldsymbol{E}$.

### Similarity Dictionary

The purpose is to add hard samples in training to make visual prompts more discriminative. As shown in the red dashed box in Fig.2, we use VLM[1] to calculate the similarity between region crop and noun phrase dictionary, and extract the top K phrase.

$$\{\boldsymbol{T_1} \dots \boldsymbol{T_K}\} = \boldsymbol{topK}\left(\boldsymbol{f^T}(\boldsymbol{T_1} \dots \boldsymbol{T_B})\right)$$

Then recalculate the similarity between K phrases, and after NMS processing, the remaining texts form a similarity dictionary. During training, negative samples by randomly sampling in the dictionary.

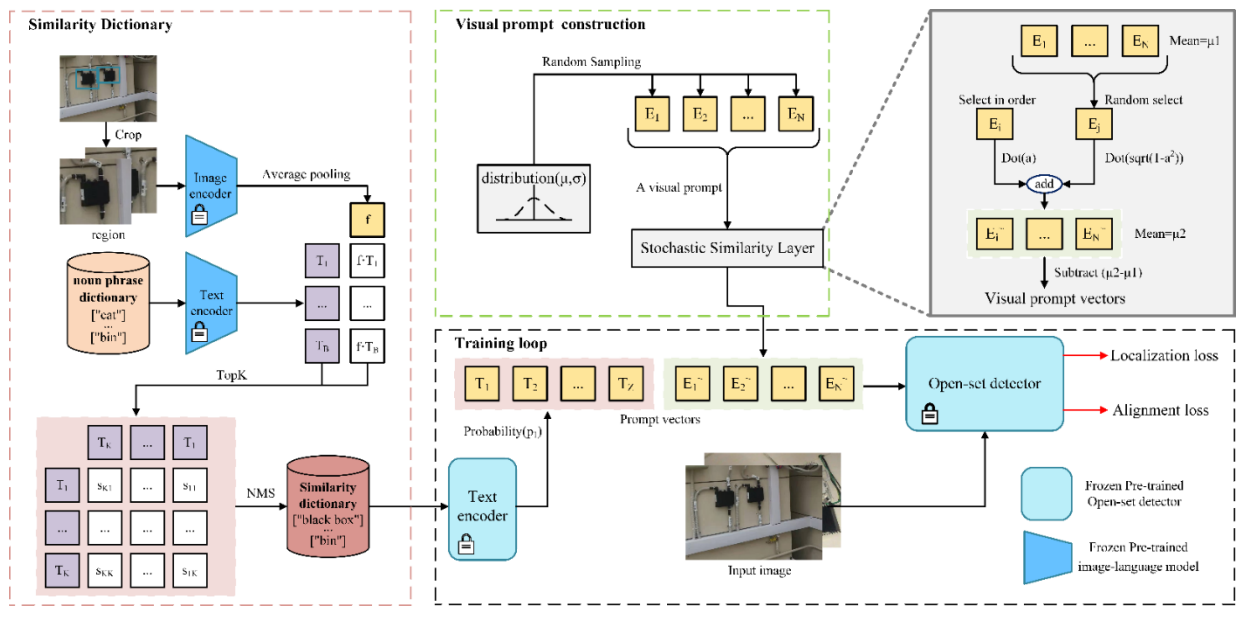



**Figure 2:** Pipeline of visual prompt training framework

## EXPERIMENTS

**Dataset:** ODinW13 is a series of real-world datasets consists of PascalVOC, Aerial Drone, Aquarium, Rabbits, Ego Hands, Mushrooms, Packages, Raccoon, Shellfish, Vehicles, Pistols, Pothole and Thermal.

**Table 1:** Comparison with other prompt methods on ODinW13

| Method | Class name initialization | Avg(mAP) |
|---|---|---|
| Text prompt | ✓ | 50.1 |
| Context Prompt[2] | ✓ | 62.1 |
| Context Prompt† | × | 57.6 |
| Offset Prompt[3] | ✓ | 64.3 |
| Offset Prompt† | × | 63 |
| Our visual prompt | **N/A** | 67.7 |

**Table 2:** The ablation study of statistical distribution, stochastic similarity layer, and similarity dictionary

| method | Statistical Distribution | Stochastic Similarity | Similarity Dictionary | mAP | mAP50 |
|---|---|---|---|---|---|
| visual prompt(A) | ✗ | ✗ | ✗ | 65.5 | 80.2 |
| visual prompt(B) | ✓ | ✗ | ✗ | 67 | 82.1 |
| visual prompt(C) | ✗ | ✓ | ✗ | 65.6 | 80.7 |
| visual prompt(D) | ✗ | ✗ | ✓ | 66.1 | 81.5 |
| visual prompt(E) | ✓ | ✓ | ✗ | 67.3 | 82.9 |
| visual prompt | ✓ | ✓ | ✓ | 67.7 | 83.5 |

**Table 3:** Explore combinatorial inference with two separately trained prompts.

| method | Similarity Dictionary | Category | mAP | mAP50 |
|---|---|---|---|---|
| visual prompt | ✗ | A | 72.6 | 79.4 |
| | ✗ | B | 51.9 | 80.8 |
| | ✗ | A+B | 54.2 | 71.6 |
| | ✓ | A | 75.6 | 81.2 |
| | ✓ | B | 52.3 | 80.8 |
| | ✓ | A+B | 60.1 | 76.7 |

### Result

In publicly datasets, our method outperforms other prompt methods. The ablation experiment verified the effectiveness of the design, and our method fully considers negative text samples, resulting in more stable performance in combination inference.

## References

**[1]** Alec Radford, Jong Wook Kim, Chris Hallacy, et al., "Learning trans ferable visual models from natural language supervision," in ICML, 2021, pp. 8748–8763.
**[2]** Chengjian Feng, Yujie Zhong, Zequn Jie, et al., "Promptdet:Towards open-vocabulary detection using uncurated im ages," in ECCV, 2022, pp. 701–717.
**[3]** Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, et al., "Grounded language-image pre-training," in CVPR, 2022, pp. 10965–10975.