# Visual Tuning

曹耘宁
2023/6/8

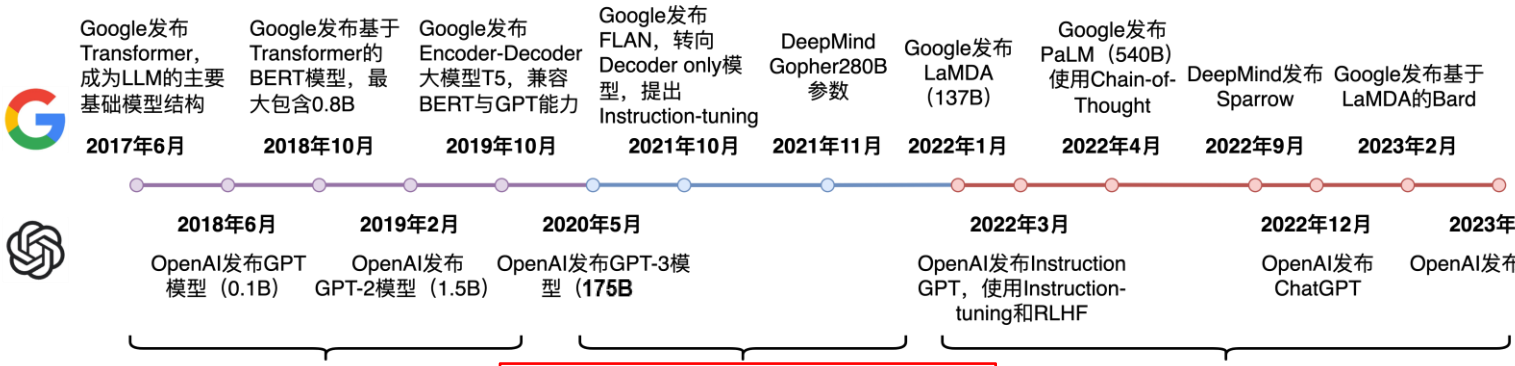
智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

- 研究背景
- Prompt tuning
- Adapter tuning
- Parameter tuning
- 总结

# 大模型的发展



- 大语言模型LLM

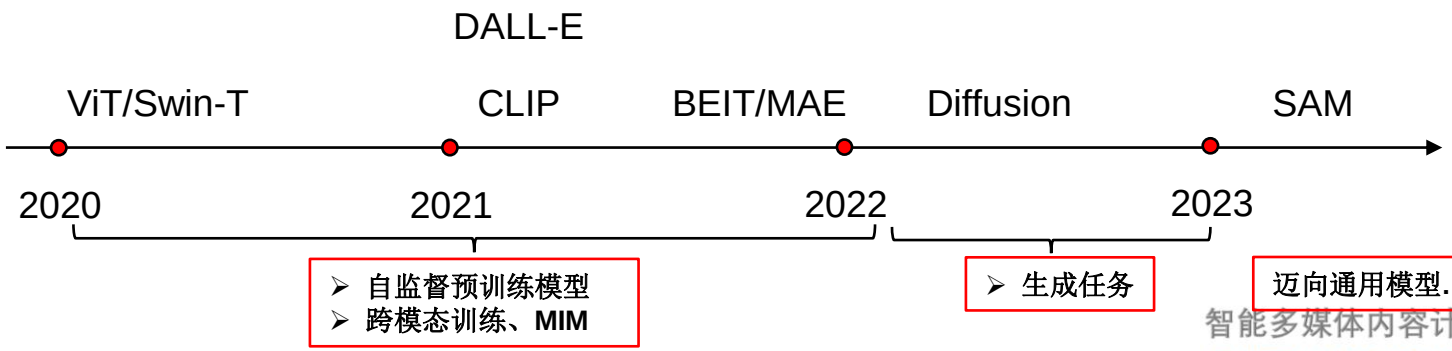



- 自监督任务的构造
- **GPT, BERT**
- **Full fine tuning**

- 生成任务占主导
- **GPT为代表的单向因果模型**
- 固定预训练模型
  **prompt tuning，参数量1%~0.1%**

- 通用模型、对话交互**(RLHF)**、模型偏见
- **GPT为代表的单向因果模型**
- 更广泛的应用**Instruction tuning，chatGPT**

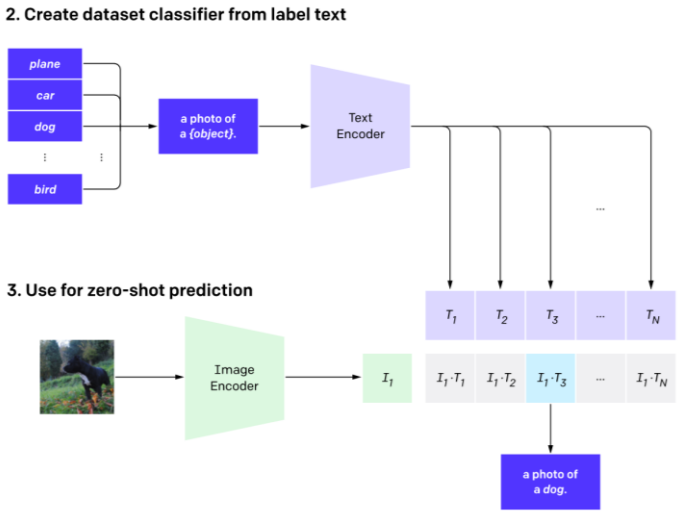
- 视觉大模型/视觉语言模型VLM



# 视觉语言模型



- CLIP，视觉语言模型
- 后续工作致力于将CLIP微调到下游任务




CLIP

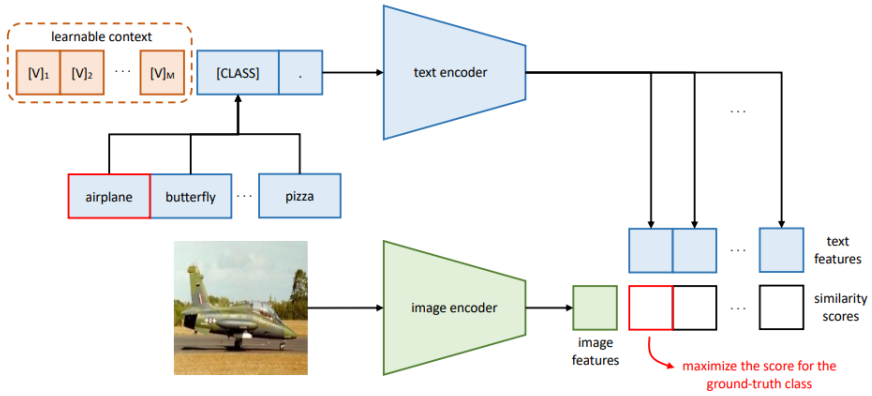



**Fig. 2 Overview of Context Optimization (CoOp).** The main idea is to model a prompt's context using a set of learnable vectors, which can be optimized through minimizing the classification loss. Two designs are proposed: one is unified context, which shares the same context vectors with all classes; and the other is class-specific context, which learns for each class a specific set of context vectors.

CoOp

# 视觉大模型



- SAM设计了适用于分割任务的多种prompt形式，实现交互式分割、多模态分割
- 近期工作使用SAM生成伪标签、对SAM做adapter tuning

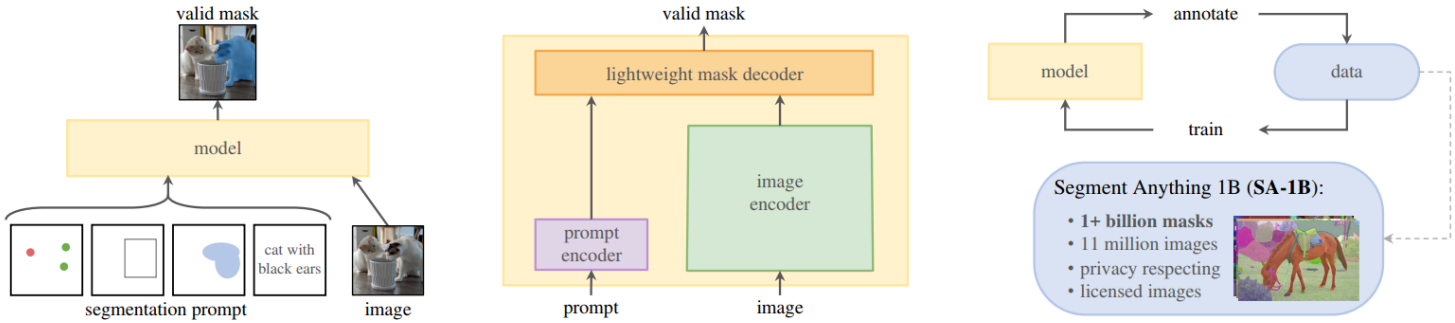



Fig. 1: Overview of the SA project, including task, model, and data. The figure is borrowed from the original paper [20].

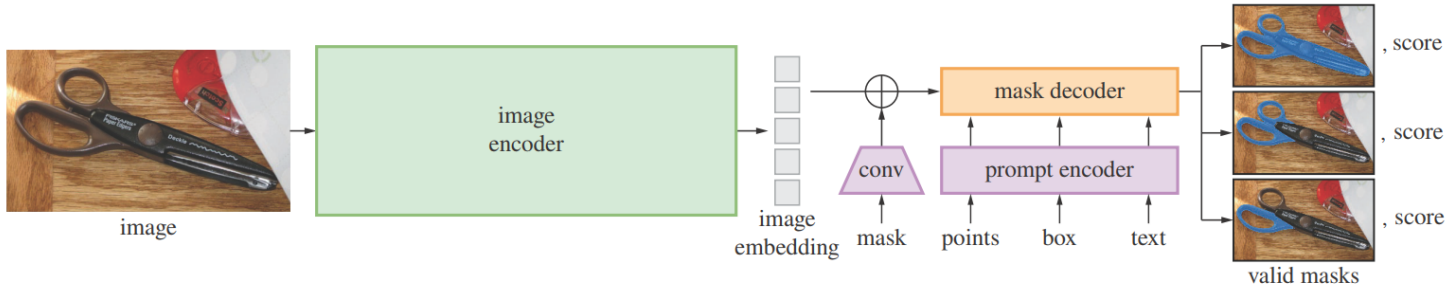
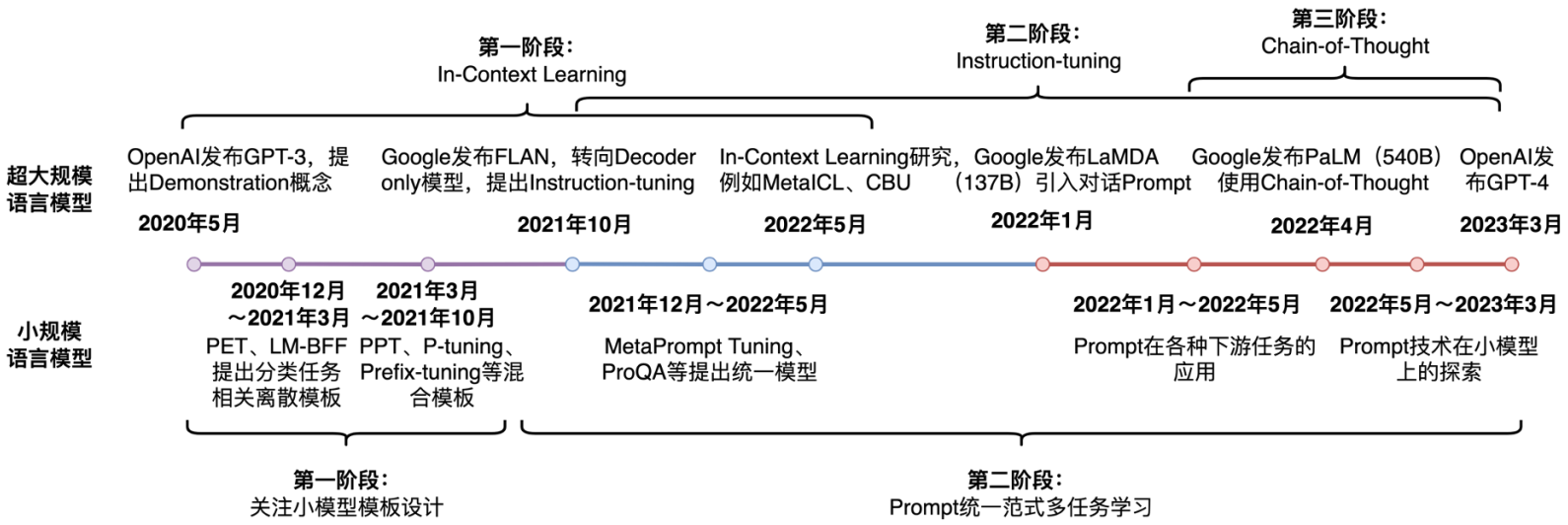



Fig. 2: Overall structure of SAM from the original paper [20].

# 语言模型微调技术

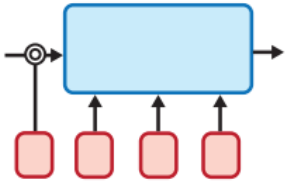
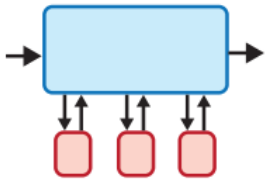
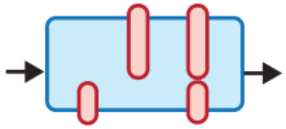


- 手工设计prompt 模板
- Prompt tuning



# 视觉微调技术分类



| Fine tuning | Prompt tuning | Adapter tuning | Parameter tuning | Remapping tuning |
|---|---|---|---|---|
|  |  |  |  |  |
| 传统方法<br>微调模型全部参数 | 继承于语言模型。冻结预训练模型，通过设计**提示模板**使预训练模型适应不同的下游任务 | 冻结预训练模型，通过向模型中插入**可训练的模块**，从而适应下游任务 | 冻结预训练模型，**直接修改其中的少量参数**，适应下游任务 | 将预训练模型的知识迁移到新模型。如知识蒸馏，NAS等 |
| 微调全量参数，计算**资源消耗太大**。当微调数据不足的时候，反而损害预训练模型的知识。 | ➢ 简单易用<br>➢ 为语言任务设计，不够灵活<br>效果与模型大小相关：大模型效果更好 | ➢ 自由插入模块，灵活方便。<br>➢ 改变了模型结构，推理负担加重 | ➢ 不改变模型结构<br>➢ 灵活度低于adapter | |

- 研究背景
- Prompt tuning
  - Visual
  - Language
  - Multi-model
- Adapter tuning
- Parameter tuning
- 总结

# Visual-driven

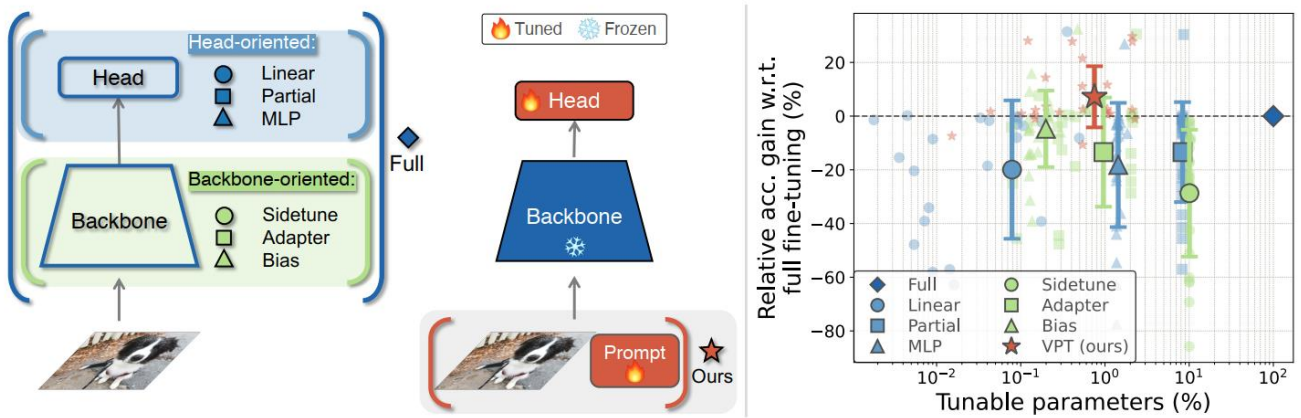


**Visual Prompt Tuning**

Menglin Jia*[1,2], Luming Tang*[1]
Bor-Chun Chen[2], Claire Cardie[1], Serge Belongie[3]
Bharath Hariharan[1], and Ser-Nam Lim[2]

[1]Cornell University [2]Meta AI [3]University of Copenhagen

- 早期工作，添加prompt把预训练ViT模型适应到不同数据集




(a) Existing tuning protocols (b) Visual-Prompt Tuning (VPT) (c) Results on visual classification tasks

**Fig. 1.** Visual-Prompt Tuning (VPT) *vs*. other transfer learning methods. (a) Current transfer learning protocols are grouped based on the tuning scope: Full fine-tuning, Head-oriented, and Backbone-oriented approaches. (b) VPT instead adds extra parameters in the input space. (c) Performance of different methods on a wide range of downstream classification tasks adapting a pre-trained ViT-B backbone, with mean and standard deviation annotated. VPT outperforms Full fine-tuning 20 out of 24 cases while using less than 1% of all model parameters

# Visual-driven



- 具体网络结构

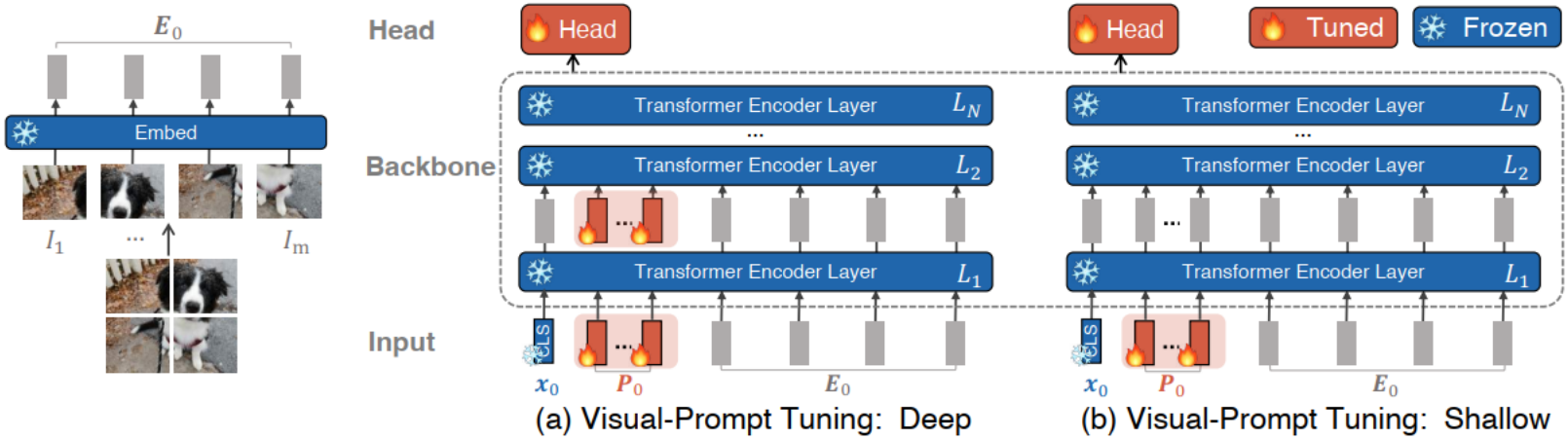



**Fig. 2.** Overview of our proposed Visual-Prompt Tuning. We explore two variants: (a) prepend a set of learnable parameters to each Transformer encoder layer's input (VPT-DEEP); (b) only insert the prompt parameters to the first layer's input (VPT-SHALLOW). During training on downstream tasks, only the parameters of prompts and linear head are updated while the whole Transformer encoder is frozen.

# Visual-driven



- 实验效果

| | ViT-B/16 (85.8M) | Total params | Scope Input | Scope Backbone | Extra params | FGVC | VTAB-1k Natural | VTAB-1k Specialized | VTAB-1k Structured |
|---|---|---|---|---|---|---|---|---|---|
| | Total # of tasks | | | | | 5 | 7 | 4 | 8 |
| (a) | FULL | 24.02× | | ✓ | | 88.54 | 75.88 | 83.36 | 47.64 |
| (b) | LINEAR | 1.02× | | | | 79.32 (0) | 68.93 (1) | 77.16 (1) | 26.84 (0) |
| | PARTIAL-1 | 3.00× | | | | 82.63 (0) | 69.44 (2) | 78.53 (0) | 34.17 (0) |
| | MLP-3 | 1.35× | | | ✓ | 79.80 (0) | 67.80 (2) | 72.83 (0) | 30.62 (0) |
| (c) | SIDETUNE | 3.69× | | ✓ | ✓ | 78.35 (0) | 58.21 (0) | 68.12 (0) | 23.41 (0) |
| | BIAS | 1.05× | | ✓ | | 88.41 (3) | 73.30 (3) | 78.25 (0) | 44.09 (2) |
| | ADAPTER | 1.23× | | ✓ | ✓ | 85.66 (2) | 70.39 (4) | 77.11 (0) | 33.43 (0) |
| (ours) | VPT-SHALLOW | 1.04× | ✓ | | ✓ | 84.62 (1) | 76.81 (4) | 79.66 (0) | 46.98 (4) |
| | VPT-DEEP | 1.18× | | | | **89.11 (4)** | **78.48 (6)** | **82.43 (2)** | **54.98 (8)** |

# Language-driven

**Texts as Images in Prompt Tuning for Multi-Label Image Recognition**

Zixian Guo[1,2*] Bowen Dong[1] Zhilong Ji[2] Jinfeng Bai[2] Yiwen Guo[4] Wangmeng Zuo[1,2(✉)]
[1]Harbin Institute of Technology [2]Tomorrow Advancing Life [3]Pazhou Lab, Guangzhou [4]Independent Researcher
zixian_guo@foxmail.com cndongsky@gmail.com zhilongji@hotmail.com
jfbai.bit@gmail.com guoyiwen89@gmail.com wmzuo@hit.edu.cn



- 早期经典方法都是language-driven，如CoOp，CoCoOp等
- CLIP Prompt tuning without image?
- 合理性在于，CLIP已经实现了视觉-语言特征对齐，(text + text encoder)与(image + image encoder)可以相互替换。当目标数据集改变，只要**把语言特征调整到目标域**，再替换image encoder即可提升性能。

文本描述
+text encoder

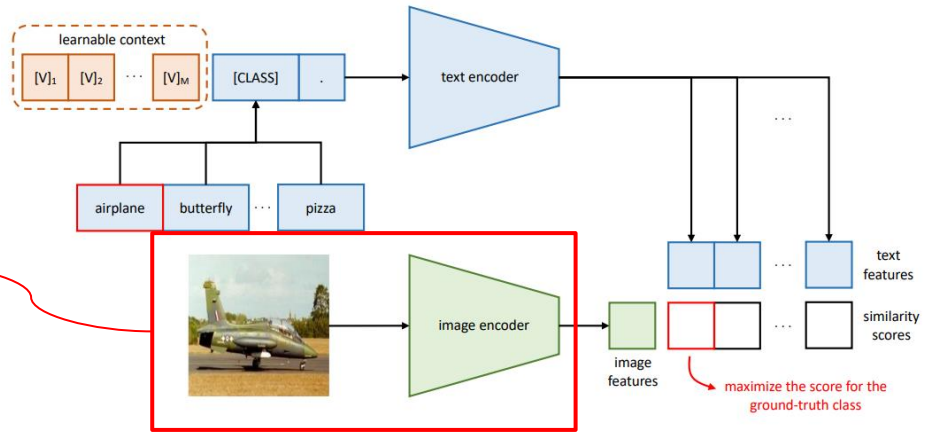



Figure 2: Overview of context optimization (CoOp).

# Language-driven



**Texts as Images in Prompt Tuning for Multi-Label Image Recognition**

Zixian Guo[1,2*] Bowen Dong[1] Zhilong Ji[2] Jinfeng Bai[2] Yiwen Guo[4] Wangmeng Zuo[1,2(✉)]
[1]Harbin Institute of Technology [2]Tomorrow Advancing Life [3]Pazhou Lab, Guangzhou [4]Independent Researcher
zixian_guo@foxmail.com cndongsky@gmail.com zhilongji@hotmail.com
jfbai.bit@gmail.com guoyiwen89@gmail.com wmzuo@hit.edu.cn

- 本文针对多标签分类（更关注标签间的结构关系），提出在训练过程不使用图像，用text description+text encoder替代image+image encoder
- 把text encoder适应到数据集标签所在的域

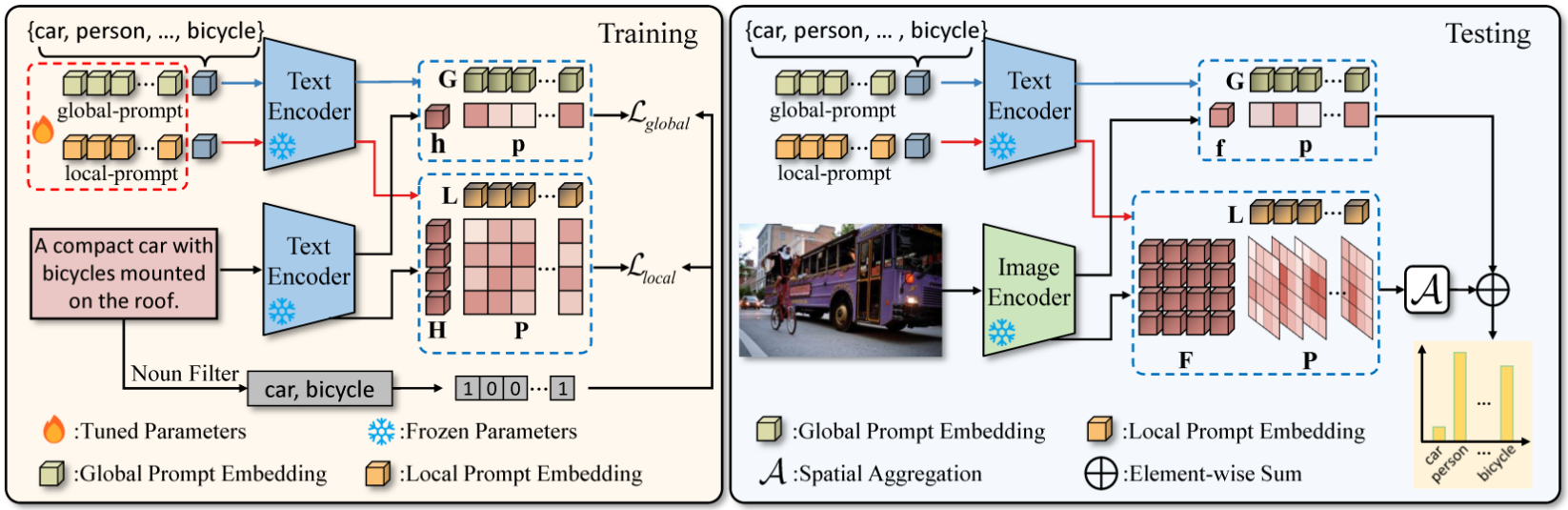



Figure 2. Training and testing pipeline of our proposed Text-as-Image (TaI) prompting, where we use text descriptions instead of labeled images to train the prompts. (a) During training, we use two identical text encoders from pre-trained CLIP to extract the global & local class embeddings (**G**&**L**) and overall & sequential text embeddings (**h**&**H**) respectively from the prompts and text description. The corresponding cosine similarity (**p**&**P**) between the embeddings are guided by the derived pseudo labels with ranking loss. (b) During testing, we replace the input from text descriptions to images. The global and local class embeddings can discriminate target classes from global & local image features (**f**&**F**). The final classification results are obtained by merging the scores of the two branches.

# Text Description



- 要求
  - 完整描述一张图的内容
  - 文本描述覆盖目标数据集的所有类别
- 直接使用MS-COCO、OpenImage等数据集的caption标注作为语料
- 类别名称的近义词词典

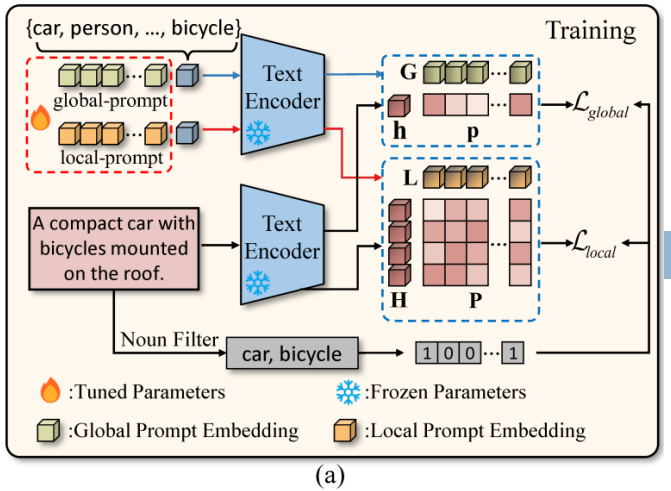
```
{'dog','pup','puppy','doggy'}
{'person','people','man','woman','human'}
{'bicycle','bike','cycle'}
{'car','taxi','automobile'}
{'boat','raft','dinghy'}
...
```

- 从语料库中筛选出包含至少一个类别名称的句子，用来训练

# Visualization



- Text prompt tuning（local）




(a)

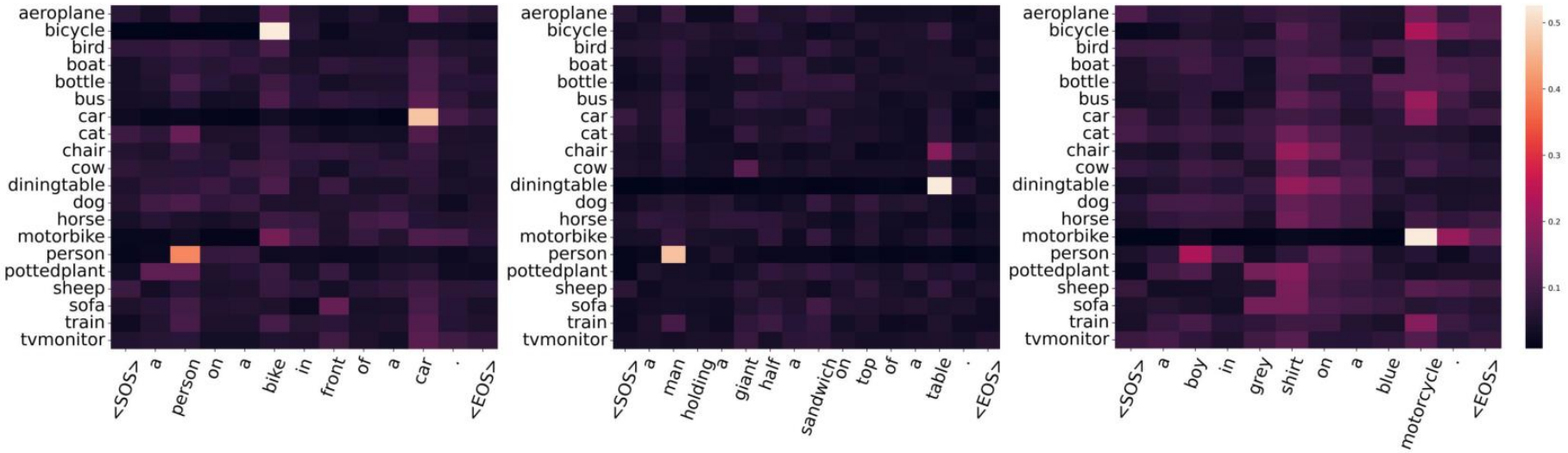



Figure 3. Visualization of correlations $\boldsymbol{P}$ between the local class embedding $\boldsymbol{L}$ and sequential token feature from texts. Each class embedding clearly correlates to words that describe the corresponding class (shown in highlight regions) rather than the global <EOS> token.

- Few-shot multi-label learning

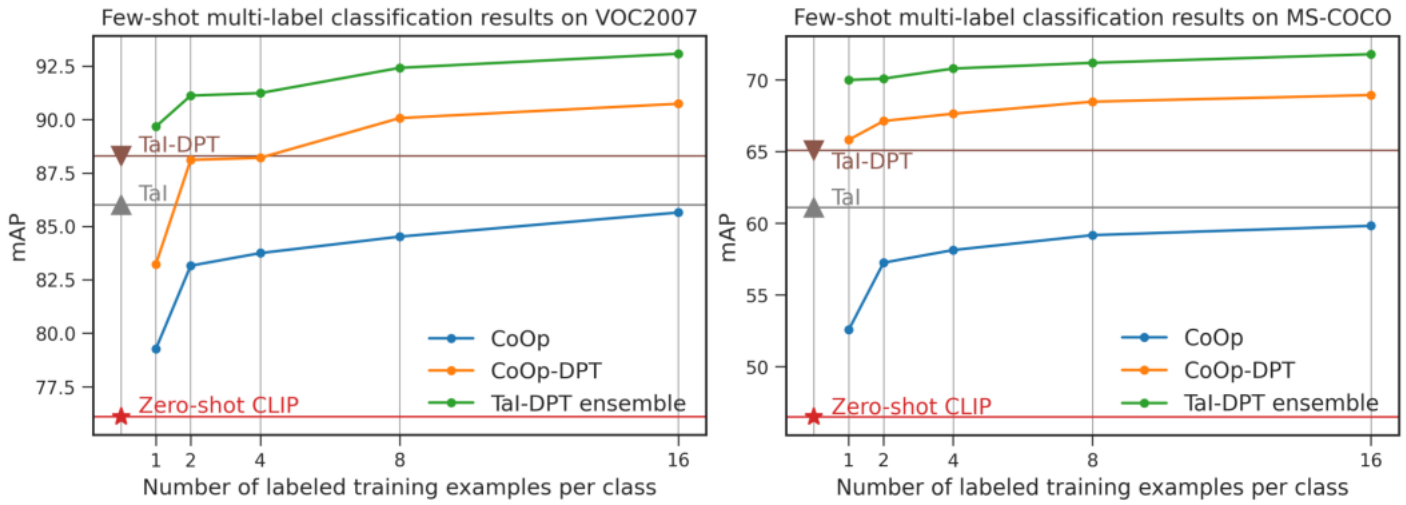



Figure 6. Comparison of different methods in few-shot multi-label recognition on VOC2007 and MS-COCO. Our zero-shot TaI-DPT can achieve comparable results with methods trained by 16-shot labeled image samples. And learned prompt ensemble proofs the complementarity between images and texts.

# Language-driven



- 性能对比

Table 1. Comparison with zero-shot methods on VOC2007, MS-COCO, and NUS-WIDE. Our proposed TaI-DPT outperforms CLIP [24] by a large margin on all datasets.

| Method | DPT | VOC2007 | MS-COCO | NUSWIDE |
|---|---|---|---|---|
| ZSCLIP | ✗ | 76.2 | 47.3 | 36.4 |
| | ✓ | 77.3 | 49.7 | 37.4 |
| TaI | ✗ | 86.0 | 61.1 | 44.9 |
| | ✓ | **88.3** | **65.1** | **46.5** |

Table 3. Results of integrating our TaI-DPT with partial-label multi-label recognition method based on pre-trained CLIP. Our approach further improves the frontier performance of DualCoOp [28]. $*$ indicates the results based on our own reproduction.

| Datasets | Method | 10% | 20% | 30% | 40% | 50% | 60% | 70% | 80% | 90% | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MS-COCO | SARB [23] | 71.2 | 75.0 | 77.1 | 78.3 | 78.9 | 79.6 | 79.8 | 80.5 | 80.5 | 77.9 |
| | DualCoOp [28] | 78.7 | 80.9 | 81.7 | 82.0 | 82.5 | 82.7 | 82.8 | 83.0 | 83.1 | 81.9 |
| | DualCoOp* | 81.0 | 82.3 | 82.9 | 83.4 | 83.5 | 83.9 | 84.0 | 84.1 | 84.3 | 83.3 |
| | +TaI-DPT | **81.5** | **82.6** | **83.3** | **83.7** | **83.9** | **84.0** | **84.2** | **84.4** | **84.5** | **83.6** |
| PascalVOC 2007 | SARB [23] | 83.5 | 88.6 | 90.7 | 91.4 | 91.9 | 92.2 | 92.6 | 92.8 | 92.9 | 90.7 |
| | DualCoOp [28] | 90.3 | 92.2 | 92.8 | 93.3 | 93.6 | 93.9 | 94.0 | 94.1 | 94.2 | 93.2 |
| | DualCoOp* | 91.4 | 93.8 | 93.8 | 94.3 | 94.6 | 94.7 | 94.8 | 94.9 | 94.9 | 94.1 |
| | +TaI-DPT | **93.3** | **94.6** | **94.8** | **94.9** | **95.1** | **95.0** | **95.1** | **95.3** | **95.5** | **94.8** |
| NUS-WIDE | DualCoOp* | 54.0 | 56.2 | 56.9 | 57.4 | 57.9 | 57.9 | 57.6 | 58.2 | 58.8 | 57.2 |
| | +TaI-DPT | **56.4** | **57.9** | **57.8** | **58.1** | **58.5** | **58.8** | **58.6** | **59.1** | **59.4** | **58.3** |

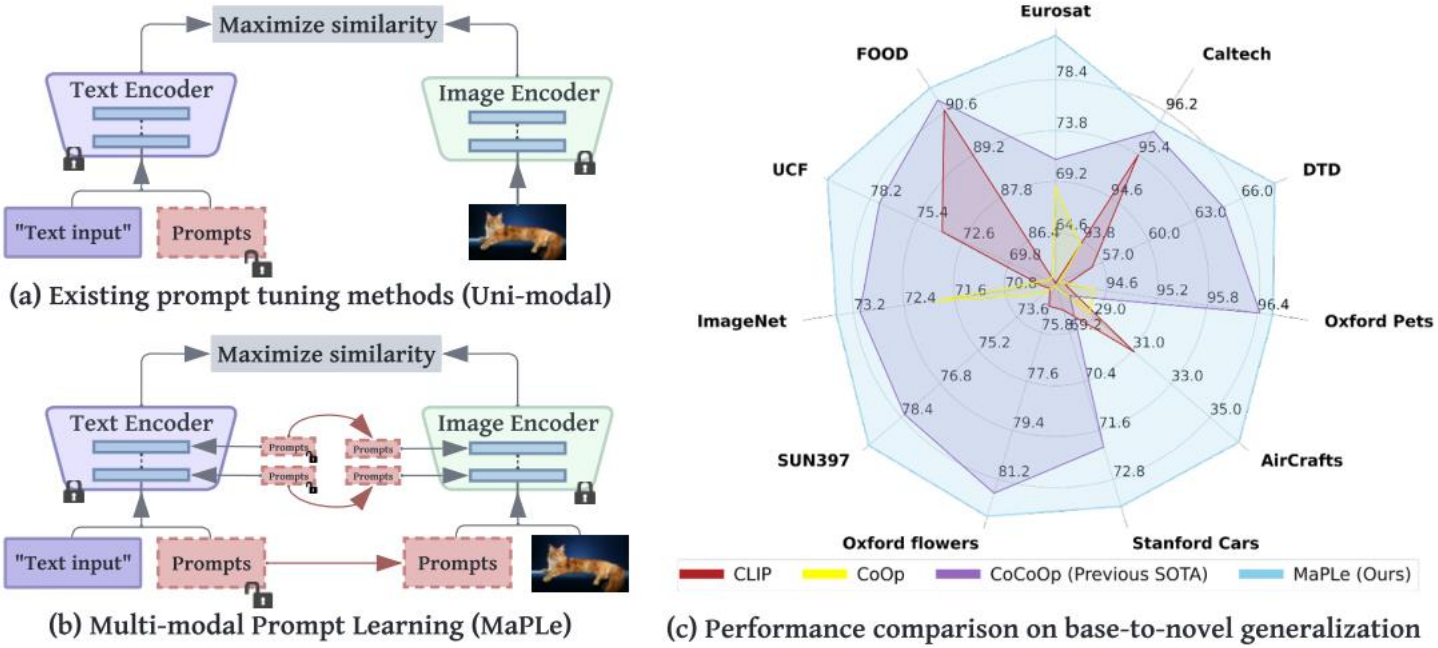
# Multi-modal

**MaPLe: Multi-modal Prompt Learning**

Muhammad Uzair Khattak[1] Hanoona Rasheed[1] Muhammad Maaz[1]
Salman Khan[1,2] Fahad Shahbaz Khan[1,3]
[1]Mohamed bin Zayed University of AI [2]Australian National University [3]Linköping University



- 仅仅对语言模型做prompt tuning有局限性
- 本文设计了语言->视觉prompt的转换器，联合调整两个模态



# Multi-modal



- Vision language prompt coupling: $\tilde{P}_k = \mathcal{F}_k(P_k)$.

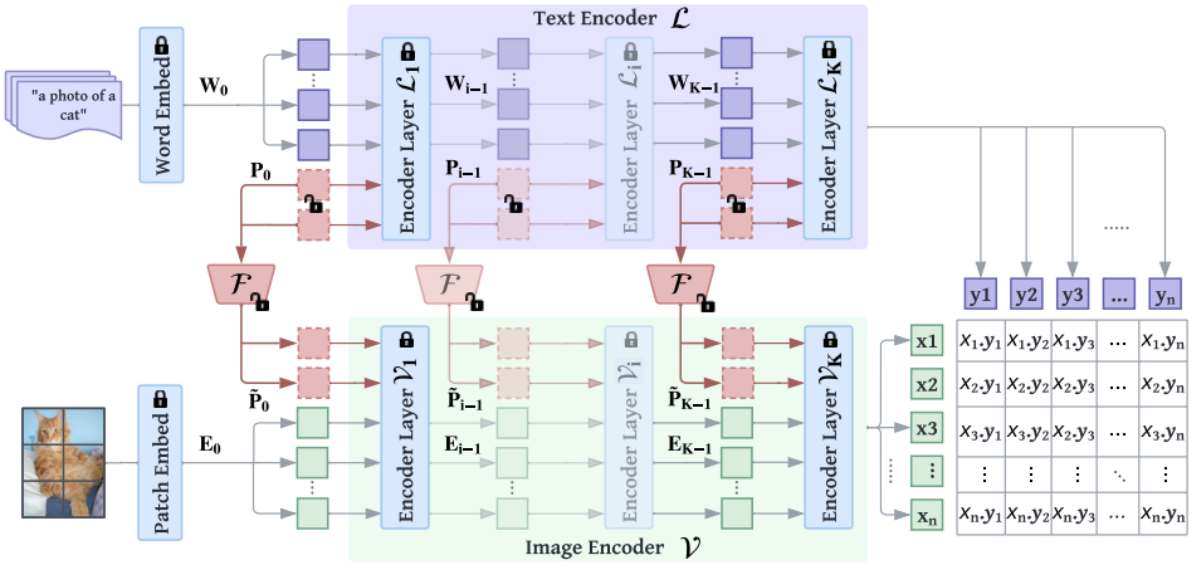



Figure 2. Overview of our proposed MaPLe (**M**ulti-moda**l** **P**rompt **Le**arning) framework for prompt learning in V-L models. MaPLe tunes both vision and language branches where only the context prompts are learned, while the rest of the model is frozen. MaPLe conditions the vision prompts on language prompts via a V-L coupling function $\mathcal{F}$ to induce mutual synergy between the two modalities. Our framework uses deep contextual prompting where separate context prompts are learned across multiple transformer blocks.

# Multi-modal



- 域泛化实验

(a) **Average over 11 datasets**

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 69.34 | 74.22 | 71.70 |
| CoOp | **82.69** | 63.22 | 71.66 |
| Co-CoOp | 80.47 | 71.69 | 75.83 |
| MaPLe | 82.28 | **75.14** | **78.55** |
| | +1.81 | +3.45 | +2.72 |

(b) ImageNet.

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 72.43 | 68.14 | 70.22 |
| CoOp | 76.47 | 67.88 | 71.92 |
| Co-CoOp | 75.98 | 70.43 | 73.10 |
| MaPLe | **76.66** | **70.54** | **73.47** |
| | +0.68 | +0.11 | +0.37 |

(c) Caltech101

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 96.84 | 94.00 | 95.40 |
| CoOp | **98.00** | 89.81 | 93.73 |
| Co-CoOp | 97.96 | 93.81 | 95.84 |
| MaPLe | 97.74 | **94.36** | **96.02** |
| | -0.22 | +0.55 | +0.18 |

(d) OxfordPets

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 91.17 | 97.26 | 94.12 |
| CoOp | 93.67 | 95.29 | 94.47 |
| Co-CoOp | 95.20 | 97.69 | 96.43 |
| MaPLe | **95.43** | **97.76** | **96.58** |
| | +0.23 | +0.07 | +0.15 |

(e) StanfordCars

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 63.37 | **74.89** | 68.65 |
| CoOp | **78.12** | 60.40 | 68.13 |
| Co-CoOp | 70.49 | 73.59 | 72.01 |
| MaPLe | 72.94 | 74.00 | **73.47** |
| | +2.45 | +0.41 | +1.46 |

(f) Flowers102

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 72.08 | **77.80** | 74.83 |
| CoOp | **97.60** | 59.67 | 74.06 |
| Co-CoOp | 94.87 | 71.75 | 81.71 |
| MaPLe | 95.92 | 72.46 | **82.56** |
| | +1.05 | +0.71 | +0.85 |

(g) Food101

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 90.10 | 91.22 | 90.66 |
| CoOp | 88.33 | 82.26 | 85.19 |
| Co-CoOp | 90.70 | 91.29 | 90.99 |
| MaPLe | **90.71** | **92.05** | **91.38** |
| | +0.01 | +0.76 | +0.39 |

(h) FGVCAircraft

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 27.19 | **36.29** | 31.09 |
| CoOp | **40.44** | 22.30 | 28.75 |
| Co-CoOp | 33.41 | 23.71 | 27.74 |
| MaPLe | 37.44 | 35.61 | **36.50** |
| | +4.03 | +11.90 | +8.76 |

(i) SUN397

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 69.36 | 75.35 | 72.23 |
| CoOp | 80.60 | 65.89 | 72.51 |
| Co-CoOp | 79.74 | 76.86 | 78.27 |
| MaPLe | **80.82** | **78.70** | **79.75** |
| | +1.08 | +1.84 | +1.48 |

(j) DTD

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 53.24 | **59.90** | 56.37 |
| CoOp | 79.44 | 41.18 | 54.24 |
| Co-CoOp | 77.01 | 56.00 | 64.85 |
| MaPLe | **80.36** | 59.18 | **68.16** |
| | +3.35 | +3.18 | +3.31 |

(k) EuroSAT

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 56.48 | 64.05 | 60.03 |
| CoOp | 92.19 | 54.74 | 68.69 |
| Co-CoOp | 87.49 | 60.04 | 71.21 |
| MaPLe | **94.07** | **73.23** | **82.35** |
| | +6.58 | +13.19 | +11.14 |

(l) UCF101

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 70.53 | 77.50 | 73.85 |
| CoOp | **84.69** | 56.05 | 67.46 |
| Co-CoOp | 82.33 | 73.45 | 77.64 |
| MaPLe | 83.00 | **78.66** | **80.77** |
| | +0.67 | +5.21 | +3.13 |

ıble 3. **Comparison with state-of-the-art methods on base-to-novel generalization**. MaPLe learns multi-modal prompts and demonrates strong generalization results over existing methods on 11 recognition datasets. Absolute gains over Co-CoOp are indicated in blue.

# 总结



- Prompt来自语言模型，本意是把不同任务转化为统一的语言描述模板输入模型，演化出各种变体。
- Prompt在语言模型效果更好，对视觉特征的改进较弱
- 因此，prompt tuning往往用于跨模态识别任务（CLIP系列）

- 检测、分割、定位等任务，需要更灵活的设计

- 研究背景
- Prompt tuning
- Adapter tuning
- Parameter tuning
- 总结

# Adapter tuning



VISION TRANSFORMER ADAPTER FOR DENSE PREDICTIONS

**Zhe Chen**[1,2*], **Yuchen Duan**[2,3*], **Wenhai Wang**[2✉], **Junjun He**[2], **Tong Lu**[1✉], **Jifeng Dai**[2,3], **Yu Qiao**[2]
[1]Nanjing University, [2]Shanghai AI Laboratory, [3]Tsinghua University
czcz94cz@gmail.com, {duanyuchen,wangwenhai,hejunjun}@pjlab.org.cn
lutong@nju.edu.cn, {daijifeng,qiaoyu}@pjlab.org.cn

- 怎样把预训练的ViT backbone适应到密集预测任务？
- 本文根据检测任务需要的先验信息设计adapter

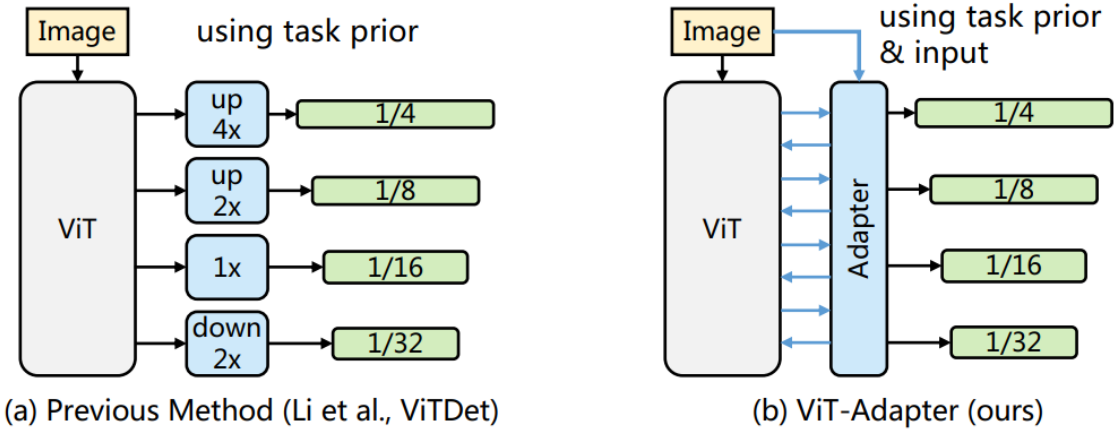



Figure 3: **Overview of ViT-Adapter and two related approaches.** Li et al. (2021b) and ViTDet (Li et al., 2022b) build simple feature pyramid to adapt plain ViT for object detection, which only consider task prior. Differently, our adapter utilizes both task prior and the input image.

# Adapter tuning



- CNN Stem，增强底层纹理信息/空间归纳偏置
- 空间特征增强
- 多尺度特征

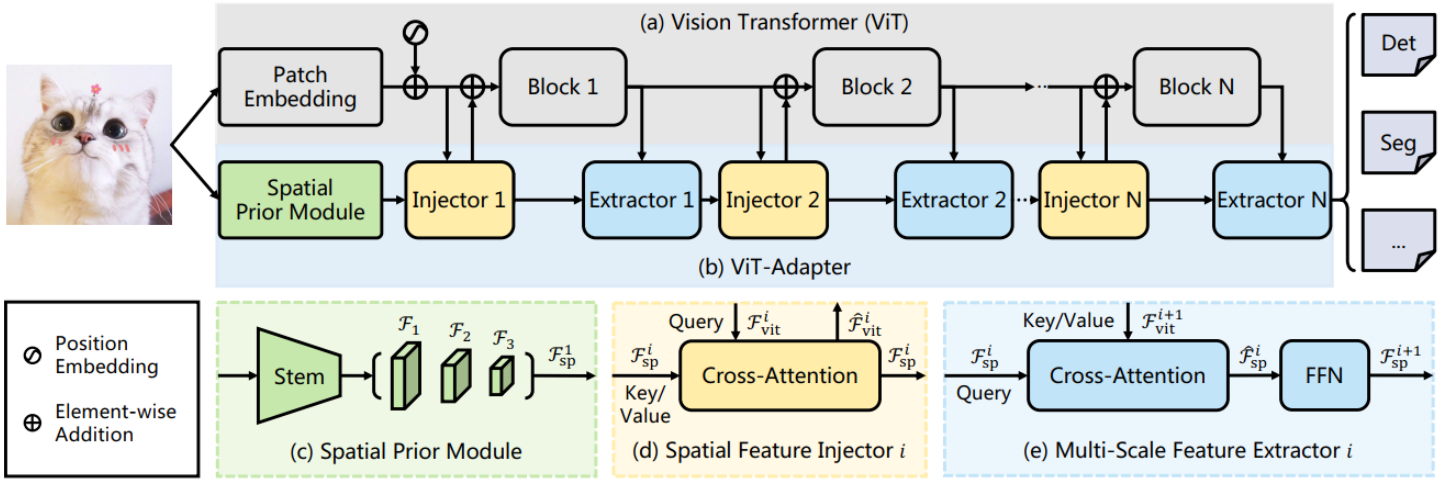
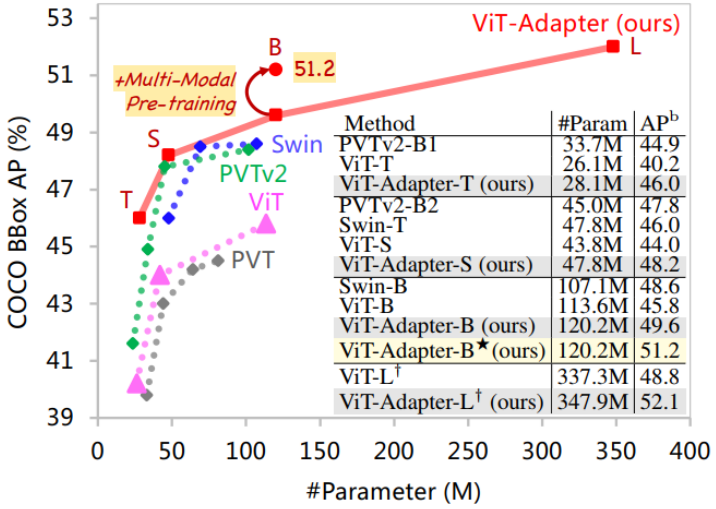



Figure 4: **Overall architecture of ViT-Adapter.** (a) The ViT, whose encoder layers are divided into $N$ (usually $N = 4$) equal blocks for feature interaction. (b) Our ViT-Adapter, which contains three key designs, including (c) a spatial prior module for modeling local spatial contexts from the input image, (d) a spatial feature injector for introducing spatial priors into ViT, and (e) a multi-scale feature extractor for reorganizing multi-scale features from the single-scale features of ViT.

# Adapter tuning



- 性能对比




| Method | #Param | $AP^b$ |
|---|---|---|
| PVTv2-B1 | 33.7M | 44.9 |
| ViT-T | 26.1M | 40.2 |
| ViT-Adapter-T (ours) | 28.1M | 46.0 |
| PVTv2-B2 | 45.0M | 47.8 |
| Swin-T | 47.8M | 46.0 |
| ViT-S | 43.8M | 44.0 |
| ViT-Adapter-S (ours) | 47.8M | 48.2 |
| Swin-B | 107.1M | 48.6 |
| ViT-B | 113.6M | 45.8 |
| ViT-Adapter-B (ours) | 120.2M | 49.6 |
| ViT-Adapter-B★ (ours) | 120.2M | 51.2 |
| ViT-L† | 337.3M | 48.8 |
| ViT-Adapter-L† (ours) | 347.9M | 52.1 |

| Method | $AP^b$ | $AP^b_{50}$ | $AP^b_{75}$ | #P |
|---|---|---|---|---|
| Cascade Mask R-CNN 3×+MS schedule | | | | |
| Swin-T (Liu et al., 2021b) | 50.5 | 69.3 | 54.9 | 86M |
| Shuffle-T (Huang et al., 2021b) | 50.8 | 69.6 | 55.1 | 86M |
| PVTv2-B2 (Wang et al., 2022a) | 51.1 | 69.8 | 55.3 | 83M |
| Focal-T (Yang et al., 2021) | 51.5 | 70.6 | 55.9 | 87M |
| ViT-S (Li et al., 2021b) | 47.9 | 67.1 | 51.7 | 82M |
| ViT-Adapter-S (ours) | 51.5 | 70.1 | 55.8 | 86M |
| Swin-B (Liu et al., 2021b) | 51.9 | 70.9 | 57.0 | 145M |
| Shuffle-B (Huang et al., 2021b) | 52.2 | 71.3 | 57.0 | 145M |
| ViT-B (Li et al., 2021b) | 50.1 | 69.3 | 54.3 | 151M |
| ViT-Adapter-B (ours) | 52.1 | 70.6 | 56.5 | 158M |

| Method | $AP^b$ | $AP^b_{50}$ | $AP^b_{75}$ | #P |
|---|---|---|---|---|
| ATSS 3×+MS schedule | | | | |
| Swin-T (Liu et al., 2021b) | 47.2 | 66.5 | 51.3 | 36M |
| Focal-T (Yang et al., 2021) | 49.5 | 68.8 | 53.9 | 37M |
| PVTv2-B2 (Wang et al., 2022a) | 49.9 | 69.1 | 54.1 | 33M |
| ViT-S (Li et al., 2021b) | 45.2 | 64.8 | 49.0 | 32M |
| ViT-Adapter-S (ours) | 49.6 | 68.5 | 54.0 | 36M |
| GFL 3×+MS schedule | | | | |
| Swin-T (Liu et al., 2021b) | 47.6 | 66.8 | 51.7 | 36M |
| PVTv2-B2 (Wang et al., 2022a) | 50.2 | 69.4 | 54.7 | 33M |
| ViT-S (Li et al., 2021b) | 46.0 | 65.5 | 49.7 | 32M |
| ViT-Adapter-S (ours) | 50.0 | 69.1 | 54.3 | 36M |

Table 2: **Object detection with different frameworks on COCO val2017.** For fair comparison, we initialize all ViT-S/B models with the regular ImageNet-1K pre-training (Touvron et al., 2021). "#P" denotes the number of parameters. "MS" means multi-scale training.

- 研究背景
- Prompt tuning
- Adapter tuning
- Parameter tuning
- 总结

# Parameter tuning



LoRA: Low-Rank Adaptation of Large Language Models

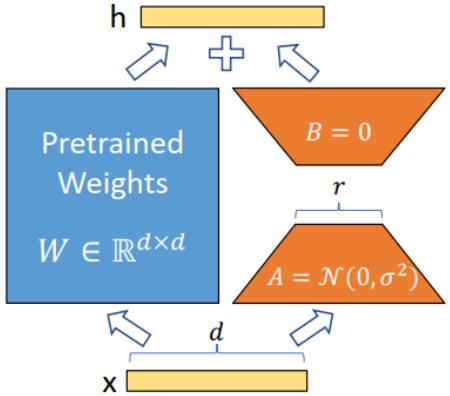
**Edward Hu*** **Yelong Shen*** **Phillip Wallis** **Zeyuan Allen-Zhu**
**Yuanzhi Li** **Shean Wang** **Lu Wang** **Weizhu Chen**
Microsoft Corporation
{edwardhu, yeshe, phwallis, zeyuana,
yuanzhil, swang, luw, wzchen}@microsoft.com
yuanzhil@andrew.cmu.edu

- Adapter改变了模型结构，推理时带来更多负担。
- 将模型参数分解为$W_0 + \Delta W$，$W_0$表示预训练参数，$\Delta W$表示微调后的变化量
- 训练阶段：$h = W_0 x + \Delta W x = W_0 x + BAx$ 呈中的变化量

- 测试阶段：将$\Delta W$加到预训练模型参数$W_0$中，不改变结构，不影响推断效率

$$W = W_0 + \Delta W$$

- 低秩近似：$\Delta W = BA,\ B \in R^{d \times r},\ A \in R^{r \times d},\ r \ll d$



# Parameter tuning



LoRA: Low-Rank Adaptation of Large Language Models

**Edward Hu*** **Yelong Shen*** **Phillip Wallis** **Zeyuan Allen-Zhu**
**Yuanzhi Li** **Shean Wang** **Lu Wang** **Weizhu Chen**
Microsoft Corporation
{edwardhu, yeshe, phwallis, zeyuana,
yuanzhil, swang, luw, wzchen}@microsoft.com
yuanzhil@andrew.cmu.edu

- 性能对比

| Model & Method | # Trainable Parameters | MNLI | SST-2 | MRPC | CoLA | QNLI | QQP | RTE | STS-B | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| $\text{RoB}_{\text{base}}$ (FT)* | 125.0M | **87.6** | 94.8 | 90.2 | **63.6** | 92.8 | **91.9** | 78.7 | 91.2 | 86.4 |
| $\text{RoB}_{\text{base}}$ (BitFit)* | 0.1M | 84.7 | 93.7 | **92.7** | 62.0 | 91.8 | 84.0 | 81.5 | 90.8 | 85.2 |
| $\text{RoB}_{\text{base}}$ ($\text{Adpt}^{\text{D}}$)* | 0.3M | $87.1_{\pm.0}$ | $94.2_{\pm.1}$ | $88.5_{\pm1.1}$ | $60.8_{\pm.4}$ | $93.1_{\pm.1}$ | $90.2_{\pm.0}$ | $71.5_{\pm2.7}$ | $89.7_{\pm.3}$ | 84.4 |
| $\text{RoB}_{\text{base}}$ ($\text{Adpt}^{\text{D}}$)* | 0.9M | $87.3_{\pm.1}$ | $94.7_{\pm.3}$ | $88.4_{\pm.1}$ | $62.6_{\pm.9}$ | $93.0_{\pm.2}$ | $90.6_{\pm.0}$ | $75.9_{\pm2.2}$ | $90.3_{\pm.1}$ | 85.4 |
| $\text{RoB}_{\text{base}}$ (LoRA) | 0.3M | $87.5_{\pm.3}$ | $\mathbf{95.1}_{\pm.2}$ | $89.7_{\pm.7}$ | $63.4_{\pm1.2}$ | $\mathbf{93.3}_{\pm.3}$ | $90.8_{\pm.1}$ | $\mathbf{86.6}_{\pm.7}$ | $\mathbf{91.5}_{\pm.2}$ | **87.2** |
| $\text{RoB}_{\text{large}}$ (FT)* | 355.0M | 90.2 | **96.4** | **90.9** | 68.0 | 94.7 | **92.2** | 86.6 | 92.4 | 88.9 |
| $\text{RoB}_{\text{large}}$ (LoRA) | 0.8M | $\mathbf{90.6}_{\pm.2}$ | $96.2_{\pm.5}$ | $\mathbf{90.9}_{\pm1.2}$ | $\mathbf{68.2}_{\pm1.9}$ | $\mathbf{94.9}_{\pm.3}$ | $91.6_{\pm.1}$ | $\mathbf{87.4}_{\pm2.5}$ | $\mathbf{92.6}_{\pm.2}$ | **89.0** |
| $\text{RoB}_{\text{large}}$ ($\text{Adpt}^{\text{P}}$)† | 3.0M | $90.2_{\pm.3}$ | $96.1_{\pm.3}$ | $90.2_{\pm.7}$ | $\mathbf{68.3}_{\pm1.0}$ | $\mathbf{94.8}_{\pm.2}$ | $\mathbf{91.9}_{\pm.1}$ | $83.8_{\pm2.9}$ | $92.1_{\pm.7}$ | 88.4 |
| $\text{RoB}_{\text{large}}$ ($\text{Adpt}^{\text{P}}$)† | 0.8M | $\mathbf{90.5}_{\pm.3}$ | $\mathbf{96.6}_{\pm.2}$ | $89.7_{\pm1.2}$ | $67.8_{\pm2.5}$ | $\mathbf{94.8}_{\pm.3}$ | $91.7_{\pm.2}$ | $80.1_{\pm2.9}$ | $91.9_{\pm.4}$ | 87.9 |
| $\text{RoB}_{\text{large}}$ ($\text{Adpt}^{\text{H}}$)† | 6.0M | $89.9_{\pm.5}$ | $96.2_{\pm.3}$ | $88.7_{\pm2.9}$ | $66.5_{\pm4.4}$ | $94.7_{\pm.2}$ | $92.1_{\pm.1}$ | $83.4_{\pm1.1}$ | $91.0_{\pm1.7}$ | 87.8 |
| $\text{RoB}_{\text{large}}$ ($\text{Adpt}^{\text{H}}$)† | 0.8M | $90.3_{\pm.3}$ | $96.3_{\pm.5}$ | $87.7_{\pm1.7}$ | $66.3_{\pm2.0}$ | $94.7_{\pm.2}$ | $91.5_{\pm.1}$ | $72.9_{\pm2.9}$ | $91.5_{\pm.5}$ | 86.4 |
| $\text{RoB}_{\text{large}}$ (LoRA)† | 0.8M | $\mathbf{90.6}_{\pm.2}$ | $96.2_{\pm.5}$ | $\mathbf{90.2}_{\pm1.0}$ | $68.2_{\pm1.9}$ | $\mathbf{94.8}_{\pm.3}$ | $91.6_{\pm.2}$ | $\mathbf{85.2}_{\pm1.1}$ | $\mathbf{92.3}_{\pm.5}$ | **88.6** |
| $\text{DeB}_{\text{XXL}}$ (FT)* | 1500.0M | 91.8 | **97.2** | 92.0 | 72.0 | **96.0** | 92.7 | 93.9 | 92.9 | 91.1 |
| $\text{DeB}_{\text{XXL}}$ (LoRA) | 4.7M | $\mathbf{91.9}_{\pm.2}$ | $96.9_{\pm.2}$ | $\mathbf{92.6}_{\pm.6}$ | $\mathbf{72.4}_{\pm1.1}$ | $\mathbf{96.0}_{\pm.1}$ | $\mathbf{92.9}_{\pm.1}$ | $\mathbf{94.9}_{\pm.4}$ | $\mathbf{93.0}_{\pm.2}$ | **91.3** |

Table 2: $\text{RoBERTa}_{\text{base}}$, $\text{RoBERTa}_{\text{large}}$, and $\text{DeBERTa}_{\text{XXL}}$ with different adaptation methods on the GLUE benchmark. We report the overall (matched and mismatched) accuracy for MNLI, Matthew's correlation for CoLA, Pearson correlation for STS-B, and accuracy for other tasks. Higher is better for all metrics. * indicates numbers published in prior works. † indicates runs configured in a setup similar to Houlsby et al. (2019) for a fair comparison.

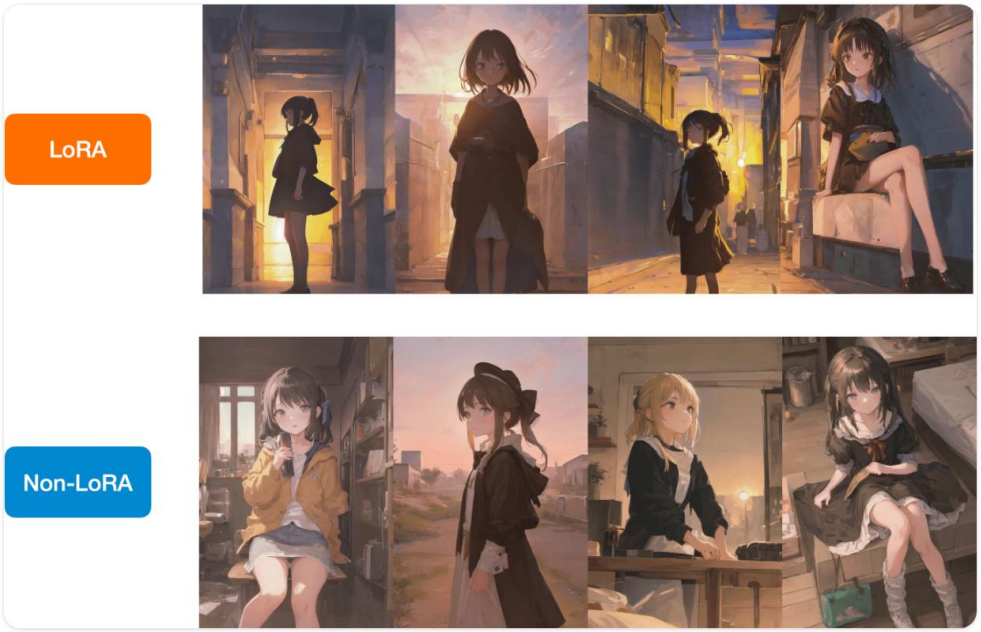
# Parameter tuning



- Parameter tuning方法不改变模型结构，更加普适。Lora最近也被广泛用于Diffusion model生成过程的微调。

prompt = "masterpiece, best quality, 1girl, at dusk"

Below is a comparison between the LoRA and the non-LoRA results:



- 研究背景
- Prompt tuning
- Adapter tuning
- Parameter tuning
- 总结

# 总结



- 大语言模型的发展整体上领先视觉模型，可以借鉴其经验
- 随着模型增大，full fine tuning会减少，由prompt，adapter等方式代替。
- 视觉模型不同于语言模型，需要更多先验知识。

  核心在于根据下游任务的特点设计相应的可微调参数


# Thanks!