JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

# Semantic Prompt for Few-Shot Image Recognition

Wentao Chen[1,2*] Chenyang Si[3*] Zhang Zhang[2,4] Liang Wang[2,4] Zilei Wang[1] Tieniu Tan[1,2,4]

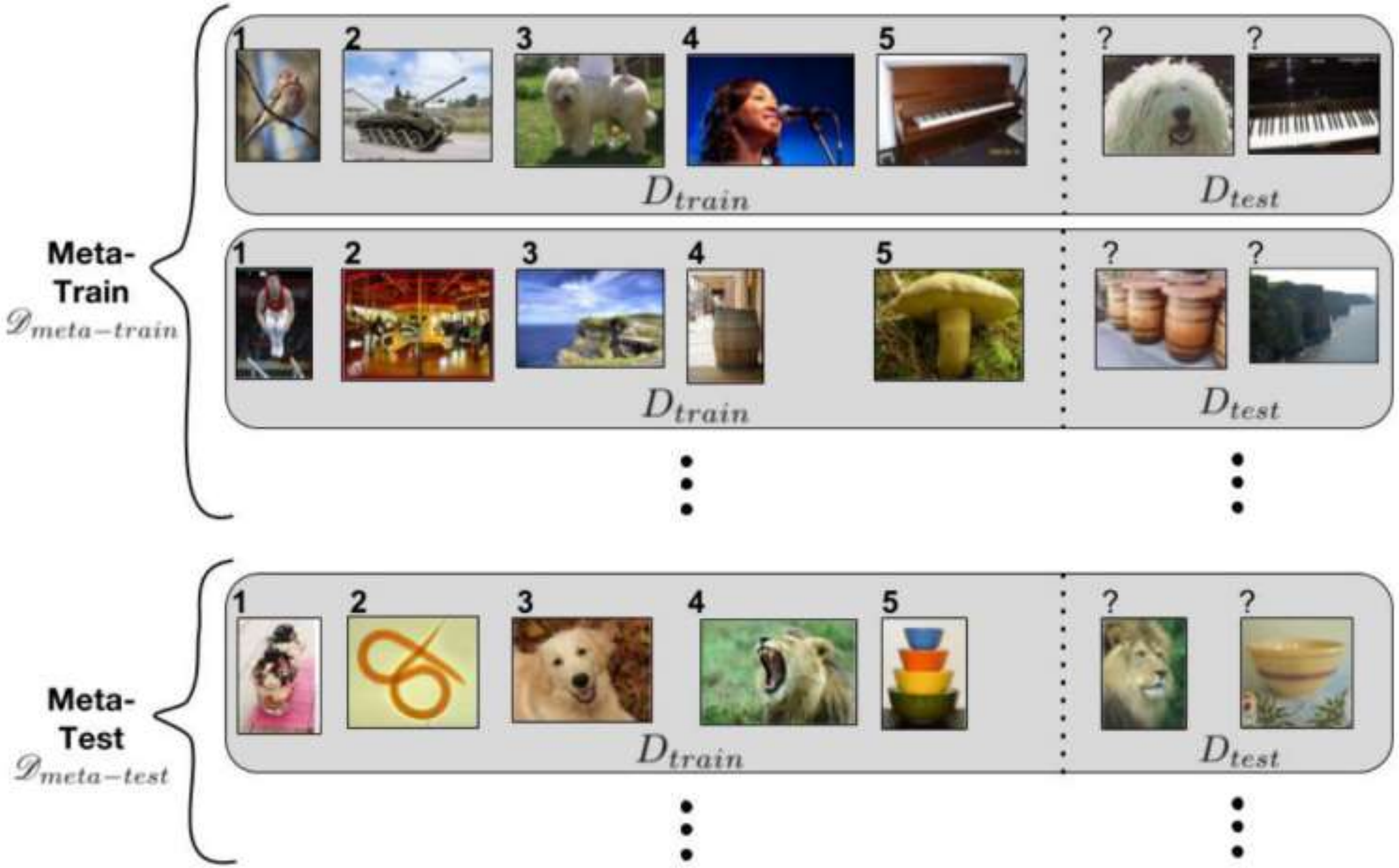
[1]University of Science and Technology of China
[2]Center for Research on Intelligent Perception and Computing, NLPR, CASIA
[3]Nanyang Technological University [4]University of Chinese Academy of Sciences

THU-PM-284

# Quick Preview

■ Task

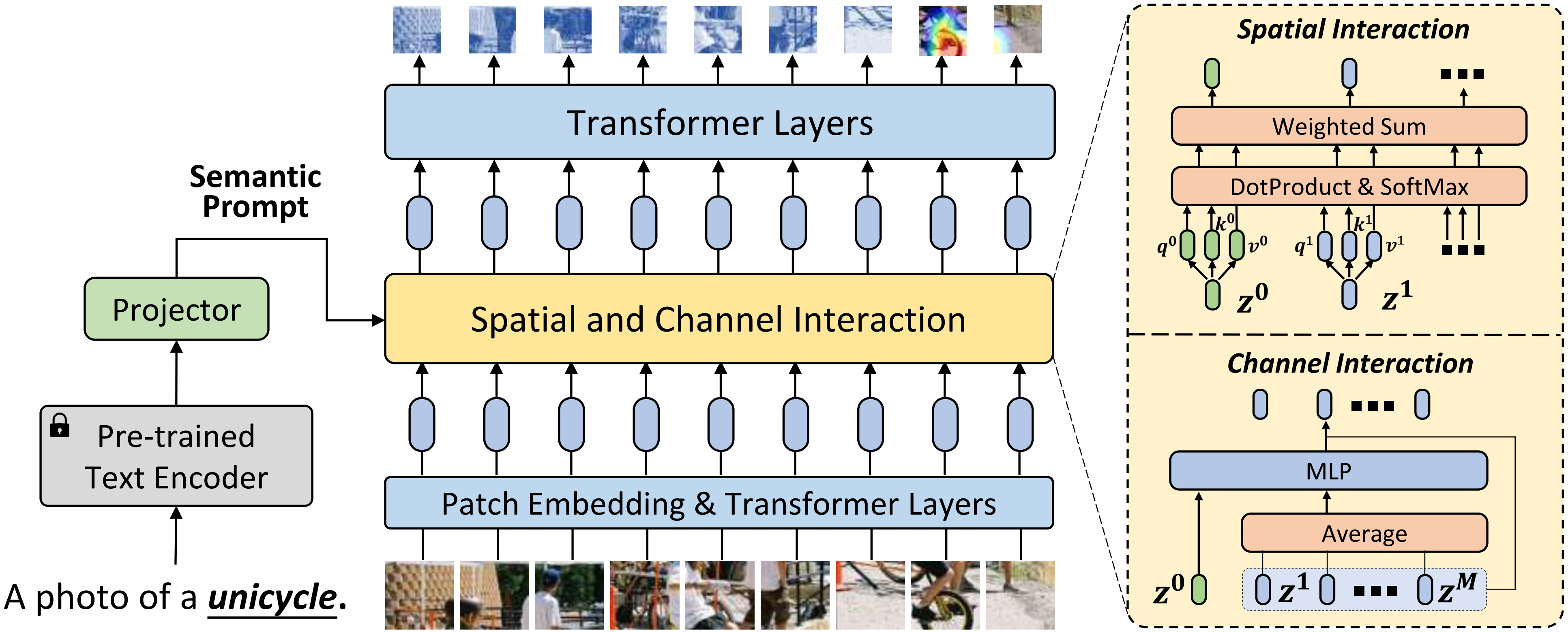
- We focus on the few-shot image recognition task, where only one or a few support images are available for a new class, and a large base dataset is used for meta-training.



# Quick Preview

## ■ Method

- We propose to use text data as semantic prompts to improve the visual feature extraction.



# Quick Preview

## ■ Experiments

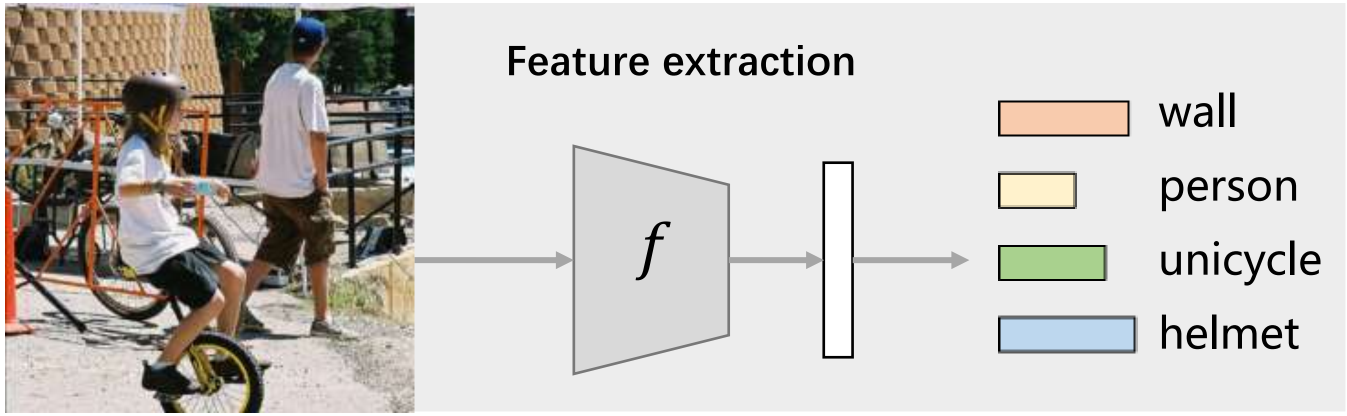
- We evluate three different text encoders, and achieve consistent improvements on four datasets.

| Method | Backbone | Params/FLOPS | *mini*ImageNet 5-way 1-shot | *mini*ImageNet 5-way 5-shot | *tiered*ImageNet 5-way 1-shot | *tiered*ImageNet 5-way 5-shot |
|---|---|---|---|---|---|---|
| LEO [42] | WRN-28-10 | 36.5M/3.7 $\times 10^{10}$ | 61.76±0.08 | 77.59±0.12 | 66.33±0.05 | 81.44±0.09 |
| CC+rot [14] | WRN-28-10 | 36.5M/3.7 $\times 10^{10}$ | 62.93±0.45 | 79.87±0.33 | 70.53±0.51 | 84.98±0.36 |
| Align [1] | WRN-28-10 | 36.5M/3.7 $\times 10^{10}$ | 65.92±0.60 | 82.85±0.55 | **74.40±0.68** | 86.61±0.59 |
| MetaOptNet [22] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 62.64±0.61 | 78.63±0.46 | 65.99±0.72 | 81.56±0.53 |
| Meta-Baseline [6] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 63.17±0.23 | 79.26±0.17 | 68.62±0.27 | 83.74±0.18 |
| DeepEMD [56] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 65.91±0.82 | 82.41±0.56 | 71.16±0.87 | 86.03±0.58 |
| RE-Net [17] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 67.60±0.44 | 82.58±0.30 | 71.61±0.51 | 85.28±0.35 |
| TPMM [51] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 67.64±0.63 | **83.44±0.43** | 72.24±0.70 | 86.55±0.63 |
| SetFeat [2] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | **68.32±0.62** | 82.71±0.46 | 73.63±0.88 | **87.59±0.57** |
| SUN [10] | Visformer-S | 12.4M/1.7 $\times 10^8$ | 67.80±0.45 | 83.25±0.30 | 72.99±0.50 | 86.74±0.33 |
| KTN [32] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 61.42±0.72 | 74.16±0.56 | - | - |
| AM3 [52] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 65.30±0.49 | 78.10±0.36 | 69.08±0.47 | 82.58±0.31 |
| TRAML [24] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | **67.10±0.52** | 79.54±0.60 | - | - |
| DeepEMD-BERT [53] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 67.03±0.79 | **83.68±0.65** | **73.76±0.72** | **87.51±0.75** |
| Pre-train (Ours) | Visformer-T | 10.0M/1.3 $\times 10^9$ | 65.16±0.44 | 81.22±0.32 | 72.38±0.50 | 86.74±0.34 |
| **SP-CLIP** (Ours) | Visformer-T | 10.0M/1.3 $\times 10^9$ | **72.31±0.40** | 83.42±0.30 | **78.03±0.46** | 88.55±0.32 |
| **SP-SBERT** (Ours) | Visformer-T | 10.0M/1.3 $\times 10^9$ | 70.70±0.42 | **83.55±0.30** | 73.31±0.50 | 88.56±0.32 |
| **SP-GloVe** (Ours) | Visformer-T | 10.0M/1.3 $\times 10^9$ | 70.81±0.42 | 83.31±0.30 | 74.68±0.50 | **88.64±0.31** |

Table 1. Comparison with previous work on *mini*ImageNet and *tiered*ImageNet. Methods in the top rows do not use semantic information, and methods in the middle rows leverage semantic information from class names [24, 32, 52] or descriptions [53]. Accuracies are reported with 95% confidence intervals.

# Motivation

- Given only one support image, the obtained image feature may contain much nioses.

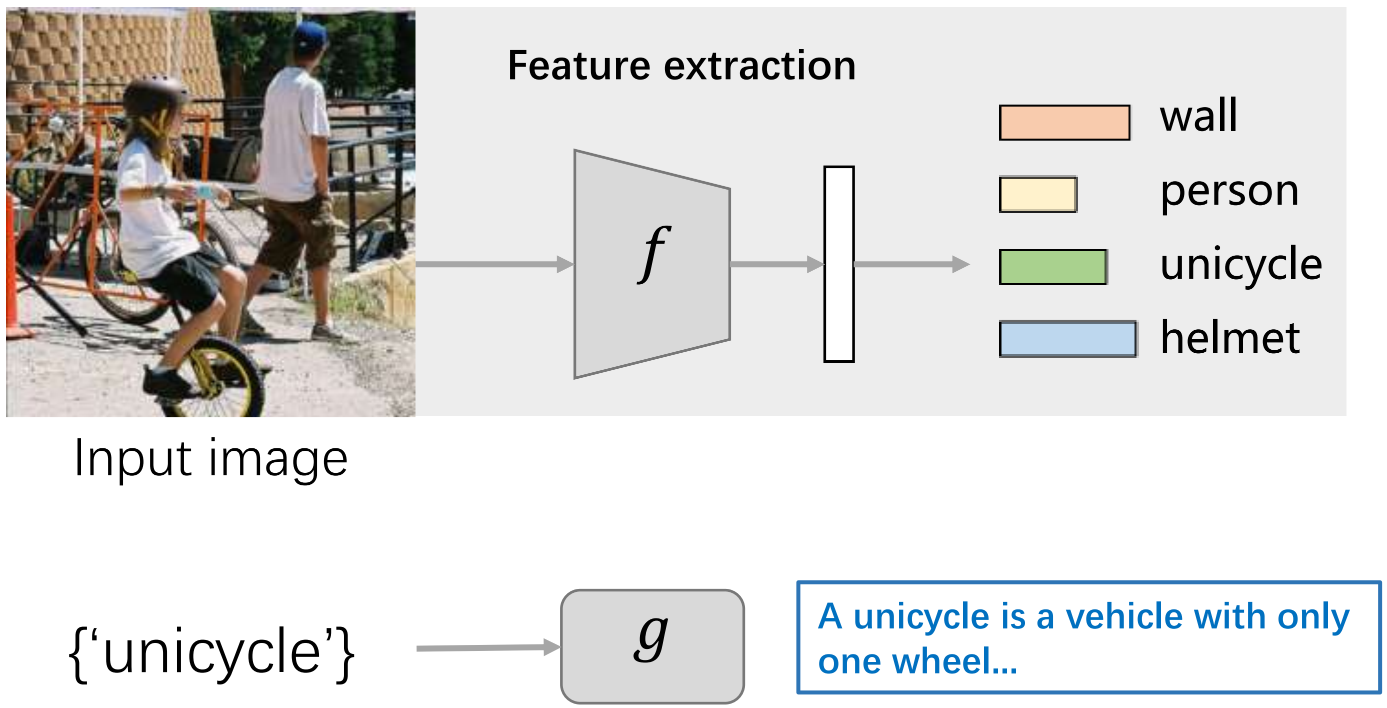



Input image

{‘unicycle’}

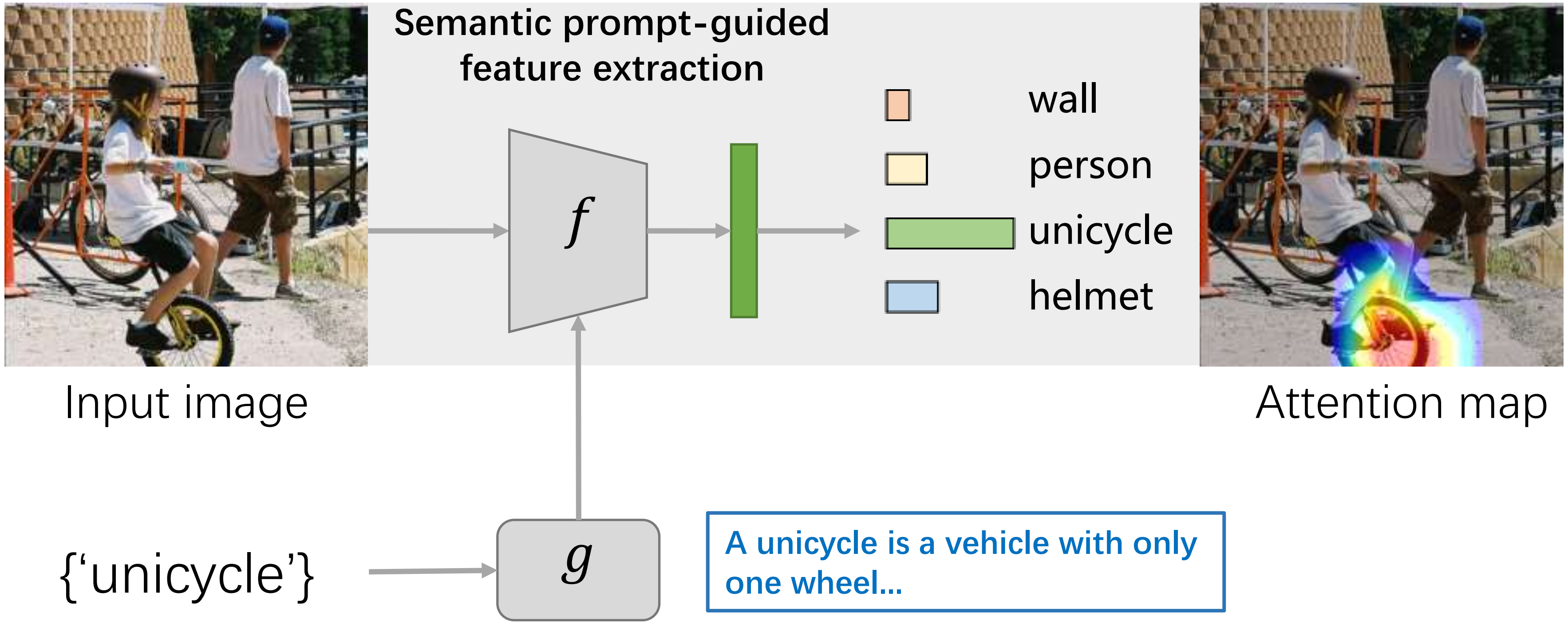
# Motivation

- Given only one support image, the obtained image feature may contain much nioses.
- The class name has rich semantic information that can be extracted by a text encoder.



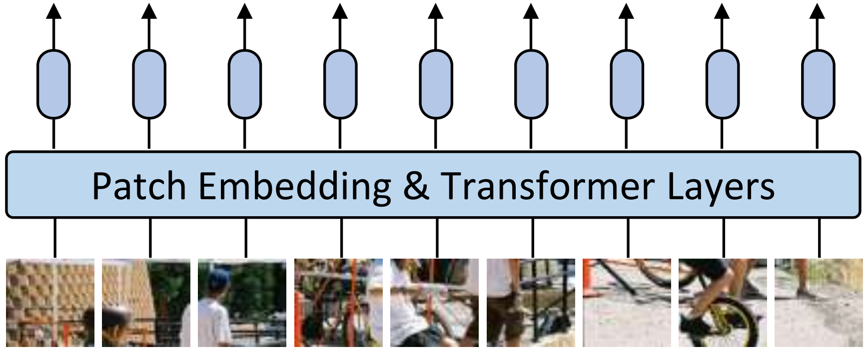
# Motivation

- Given only one support image, the obtained image feature may contain much nioses.
- The class name has rich semantic information that can be extracted by a text encoder.
- We use semantic features as prompts to improve the visual feature extraction.



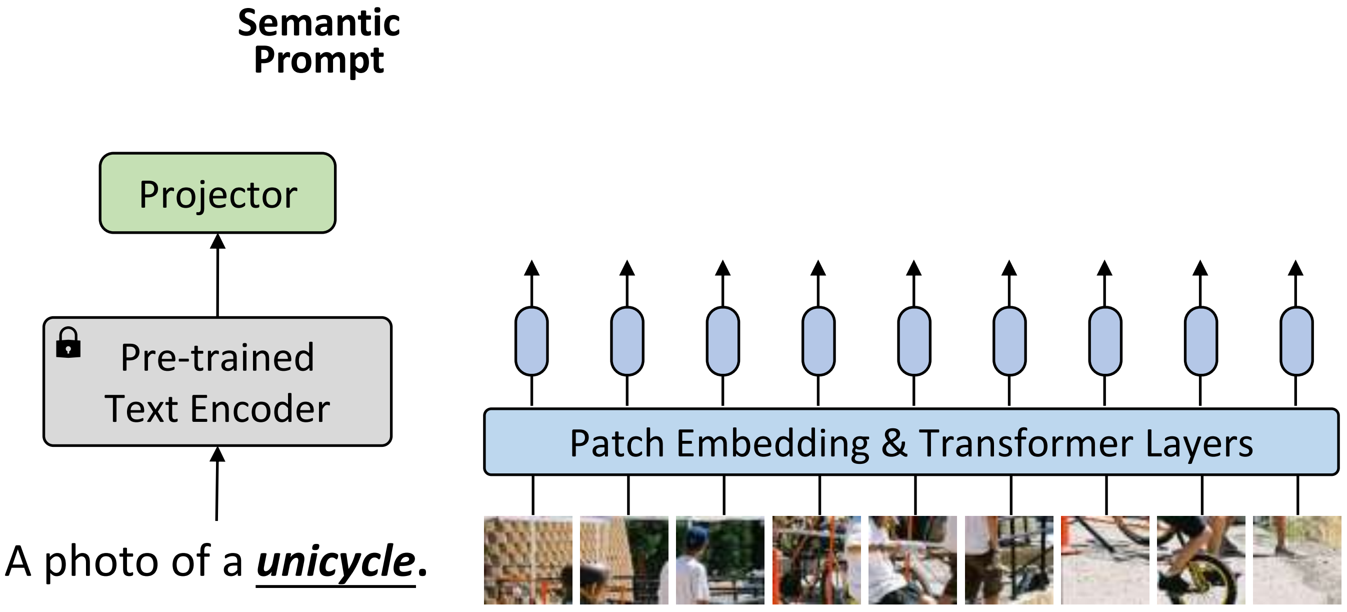
# The framework of semantic prompt

- Feed image patches into a Vision Transformer.



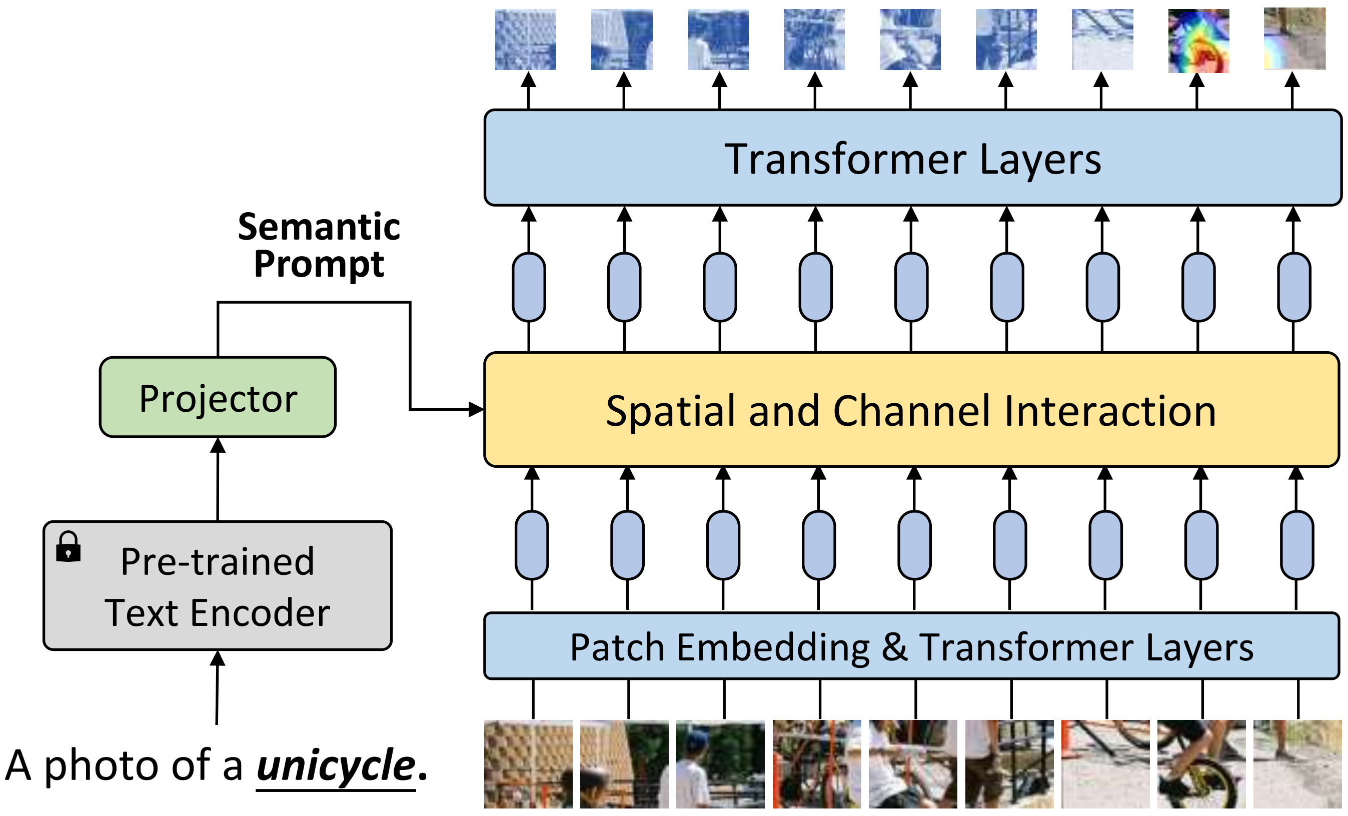
# The framework of semantic prompt

- Feed image patches into a Vision Transformer.
- Feed the class name into a text encoder to obtain a semantic prompt.



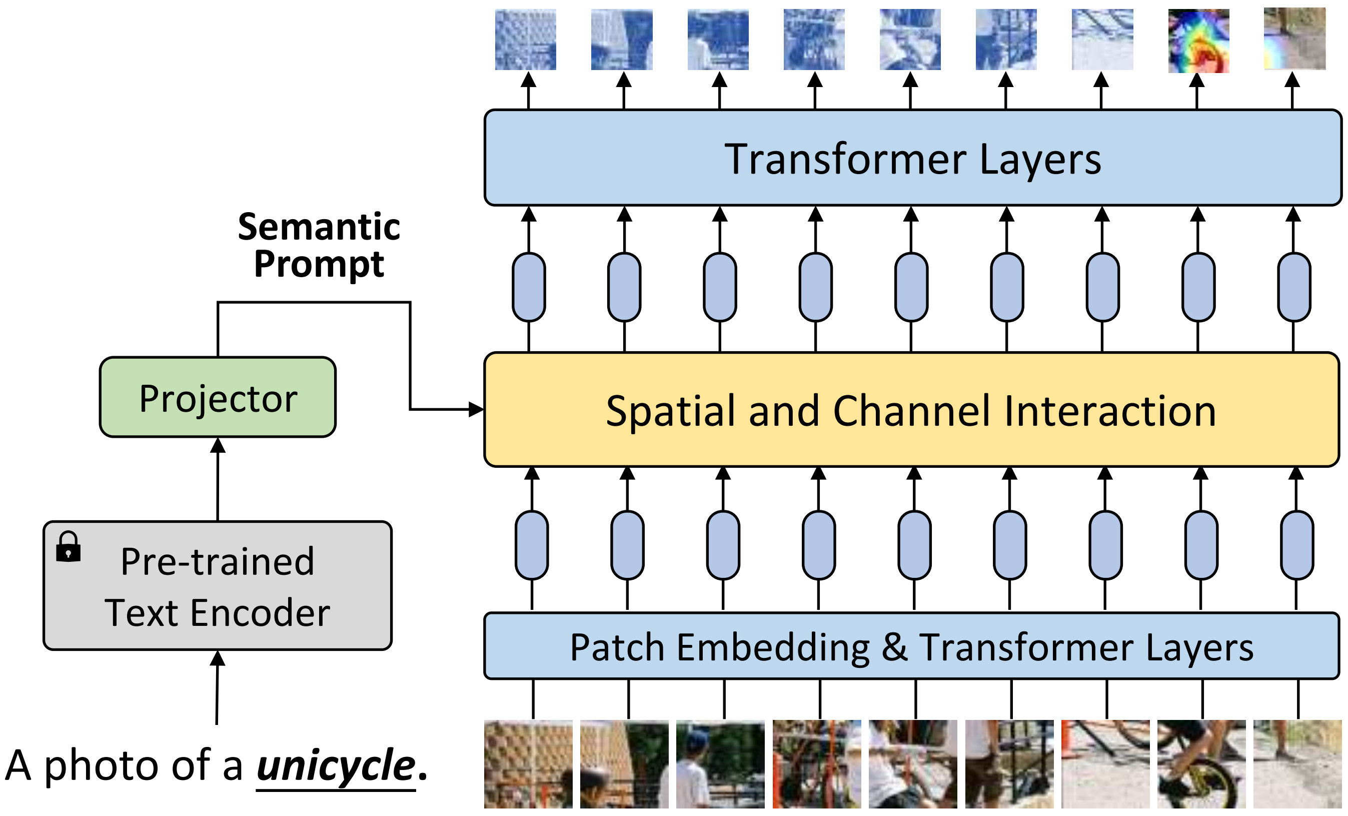
# The framework of semantic prompt

- Feed image patches into a Vision Transformer.
- Feed the class name into a text encoder to obtain a semantic prompt.
- Extract image features guided by the semantic prompt via spatial and channel interaction.



# The framework of semantic prompt

- Feed image patches into a Vision Transformer.
- Feed the class name into a text encoder to obtain a semantic prompt.
- Extract image features guided by the semantic prompt via spatial and channel interaction.
- Train the model via meta-learning.




1.class prototype:

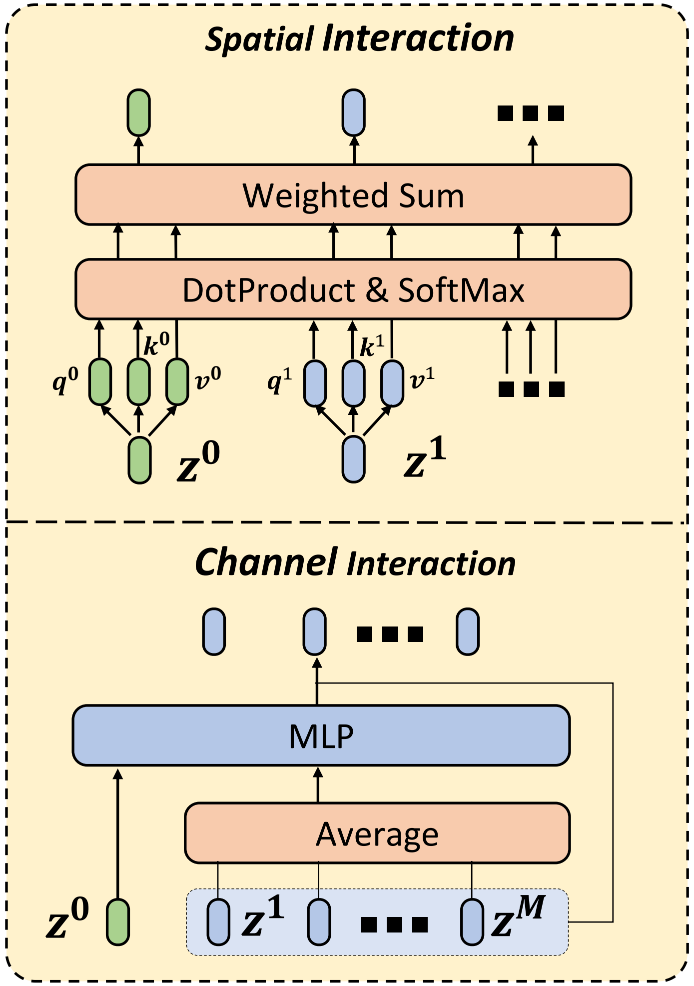
$$c_i = \frac{1}{K}\sum_{j=1}^{K} f_g(x_j^s)$$

2.loss function:

$$L_{meta} = -\mathbb{E}_{S,Q}\mathbb{E}_{x^q} \log \frac{\exp(s(f(x^q), c_{y^q})/\tau)}{\sum_{i=1}^{N} \exp(s(f(x^q), c_i)/\tau)}$$

# Spatial and channel interaction

- Adapt visual features on spatial and channel dimensions according to the given prompt.




- Spatial Interaction

  ① Concat the prompt and patches.

  $$\widehat{\boldsymbol{Z}}_{l-1} = [\boldsymbol{z}^0, \boldsymbol{z}_{l-1}^1, \ldots, \boldsymbol{z}_{l-1}^M]$$

  ② Interact with multi-head attention.

  $$[\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}] = \widehat{\boldsymbol{Z}}_{l-1}\boldsymbol{W}_{qkv}$$

  $$\boldsymbol{A} = softmax(\boldsymbol{q}\boldsymbol{k}^T / C_h^{1/4})$$

  $$MSA(\widehat{\boldsymbol{Z}}_{l-1}) = (\boldsymbol{A}v)\boldsymbol{W}_{out}$$

- Channel Interaction

  ① Average patch features: $\boldsymbol{z}_{l-1}^c = \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{z}_{l-1}^i$

  ② Feed the prompt and visual context into MLP.

  $$\boldsymbol{\beta}_{l-1} = MLP([\boldsymbol{z}^0; \boldsymbol{z}_{l-1}^c])$$

  ③ Add the bias vector to all patch features.

  $$\widehat{\boldsymbol{Z}}_{l-1} = [\boldsymbol{z}_{l-1}^i + \boldsymbol{\beta}_{l-1},] \quad i = 1,2, \ldots, M$$

# Experimental results

- miniImageNet & tieredImageNet

| Method | Backbone | Params/FLOPS | *mini*ImageNet 5-way 1-shot | *mini*ImageNet 5-way 5-shot | *tiered*ImageNet 5-way 1-shot | *tiered*ImageNet 5-way 5-shot |
|---|---|---|---|---|---|---|
| LEO [42] | WRN-28-10 | $36.5M/3.7 \times 10^{10}$ | 61.76±0.08 | 77.59±0.12 | 66.33±0.05 | 81.44±0.09 |
| CC+rot [14] | WRN-28-10 | $36.5M/3.7 \times 10^{10}$ | 62.93±0.45 | 79.87±0.33 | 70.53±0.51 | 84.98±0.36 |
| Align [1] | WRN-28-10 | $36.5M/3.7 \times 10^{10}$ | 65.92±0.60 | 82.85±0.55 | **74.40±0.68** | 86.61±0.59 |
| MetaOptNet [22] | ResNet-12 | $12.5M/3.5 \times 10^{9}$ | 62.64±0.61 | 78.63±0.46 | 65.99±0.72 | 81.56±0.53 |
| Meta-Baseline [6] | ResNet-12 | $12.5M/3.5 \times 10^{9}$ | 63.17±0.23 | 79.26±0.17 | 68.62±0.27 | 83.74±0.18 |
| DeepEMD [56] | ResNet-12 | $12.5M/3.5 \times 10^{9}$ | 65.91±0.82 | 82.41±0.56 | 71.16±0.87 | 86.03±0.58 |
| RE-Net [17] | ResNet-12 | $12.5M/3.5 \times 10^{9}$ | 67.60±0.44 | 82.58±0.30 | 71.61±0.51 | 85.28±0.35 |
| TPMM [51] | ResNet-12 | $12.5M/3.5 \times 10^{9}$ | 67.64±0.63 | **83.44±0.43** | 72.24±0.70 | 86.55±0.63 |
| SetFeat [2] | ResNet-12 | $12.5M/3.5 \times 10^{9}$ | **68.32±0.62** | 82.71±0.46 | 73.63±0.88 | **87.59±0.57** |
| SUN [10] | Visformer-S | $12.4M/1.7 \times 10^{8}$ | 67.80±0.45 | 83.25±0.30 | 72.99±0.50 | 86.74±0.33 |
| KTN [32] | ResNet-12 | $12.5M/3.5 \times 10^{9}$ | 61.42±0.72 | 74.16±0.56 | - | - |
| AM3 [52] | ResNet-12 | $12.5M/3.5 \times 10^{9}$ | 65.30±0.49 | 78.10±0.36 | 69.08±0.47 | 82.58±0.31 |
| TRAML [24] | ResNet-12 | $12.5M/3.5 \times 10^{9}$ | **67.10±0.52** | 79.54±0.60 | - | - |
| DeepEMD-BERT [53] | ResNet-12 | $12.5M/3.5 \times 10^{9}$ | 67.03±0.79 | **83.68±0.65** | **73.76±0.72** | **87.51±0.75** |
| Pre-train (Ours) | Visformer-T | $10.0M/1.3 \times 10^{9}$ | 65.16±0.44 | 81.22±0.32 | 72.38±0.50 | 86.74±0.34 |
| **SP-CLIP** (Ours) | Visformer-T | $10.0M/1.3 \times 10^{9}$ | **72.31±0.40** | 83.42±0.30 | **78.03±0.46** | 88.55±0.32 |
| **SP-SBERT** (Ours) | Visformer-T | $10.0M/1.3 \times 10^{9}$ | 70.70±0.42 | **83.55±0.30** | 73.31±0.50 | 88.56±0.32 |
| **SP-GloVe** (Ours) | Visformer-T | $10.0M/1.3 \times 10^{9}$ | 70.81±0.42 | 83.31±0.30 | 74.68±0.50 | **88.64±0.31** |

Table 1. Comparison with previous work on *mini*ImageNet and *tiered*ImageNet. Methods in the top rows do not use semantic information, and methods in the middle rows leverage semantic information from class names [24, 32, 52] or descriptions [53]. Accuracies are reported with 95% confidence intervals.

# Experimental results

- CIFAR-FS & FC100

| Method | Backbone | Params/FLOPs | CIFAR-FS 5-way 1-shot | CIFAR-FS 5-way 5-shot | FC100 5-way 1-shot | FC100 5-way 5-shot |
|---|---|---|---|---|---|---|
| PN+rot [14] | WRN-28-10 | 36.5M/3.7 $\times 10^{10}$ | 69.55±0.34 | 82.34±0.24 | - | - |
| Align [1] | WRN-28-10 | 36.5M/3.7 $\times 10^{10}$ | - | - | **45.83±0.48** | 59.74±0.56 |
| ProtoNet [45] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 72.2±0.7 | 83.5±0.5 | 37.5±0.6 | 52.5±0.6 |
| MetaOptNet [22] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 72.6±0.7 | 84.3±0.5 | 41.1±0.6 | 55.5±0.6 |
| MABAS [18] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 73.51±0.92 | 85.49±0.68 | 42.31±0.75 | 57.56±0.78 |
| Distill [47] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 73.9±0.8 | 86.9±0.5 | 44.6±0.7 | **60.9±0.6** |
| RE-Net [17] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | 74.51±0.46 | 86.60±0.32 | - | - |
| infoPatch [27] | ResNet-12 | 12.5M/3.5 $\times 10^9$ | - | - | 43.8±0.4 | 58.0±0.4 |
| SUN [10] | Visformer-S | 12.4M/1.7 $\times 10^8$ | **78.37±0.46** | **88.84±0.32** | - | - |
| Pre-train (Ours) | Visformer-T | 10.0M/1.3 $\times 10^9$ | 71.99±0.47 | 85.98±0.34 | 43.77±0.39 | 59.48±0.39 |
| **SP-CLIP** (Ours) | Visformer-T | 10.0M/1.3 $\times 10^9$ | **82.18±0.40** | 88.24±0.32 | **48.53±0.38** | **61.55±0.41** |
| **SP-SBERT** (Ours) | Visformer-T | 10.0M/1.3 $\times 10^9$ | 81.32±0.40 | 88.31±0.32 | 47.03±0.40 | 61.03±0.40 |
| **SP-GloVe** (Ours) | Visformer-T | 10.0M/1.3 $\times 10^9$ | 81.62±0.41 | **88.32±0.32** | 46.69±0.41 | 61.18±0.41 |

Table 2. Comparison with previous work on CIFAR-FS [22] and FC100 [31].

# Experimental results

| | Aug | SI | CI | Mini | Tiered | CIFAR-FS | FC100 |
|---|---|---|---|---|---|---|---|
| | × | × | × | 61.96 | 71.91 | 68.84 | 40.78 |
| | ✓ | × | × | 65.15 | 72.38 | 71.99 | 43.77 |
| ↑ 5.9% | ✓ | ✓ | × | 71.59 | 76.20 | 81.19 | 47.83 |
| ↑ 5.4% | ✓ | × | ✓ | 70.48 | 77.62 | 79.80 | 47.10 |
| ↑ 6.9% | ✓ | ✓ | ✓ | **72.31** | **78.03** | **82.18** | **48.53** |

Table 3. Ablation study on four datasets under the 1-shot setting. SI means spatial interaction, and CI means channel interaction.


Figure 4. Visualization of attention maps when prompting with different class labels.

# Experimental results

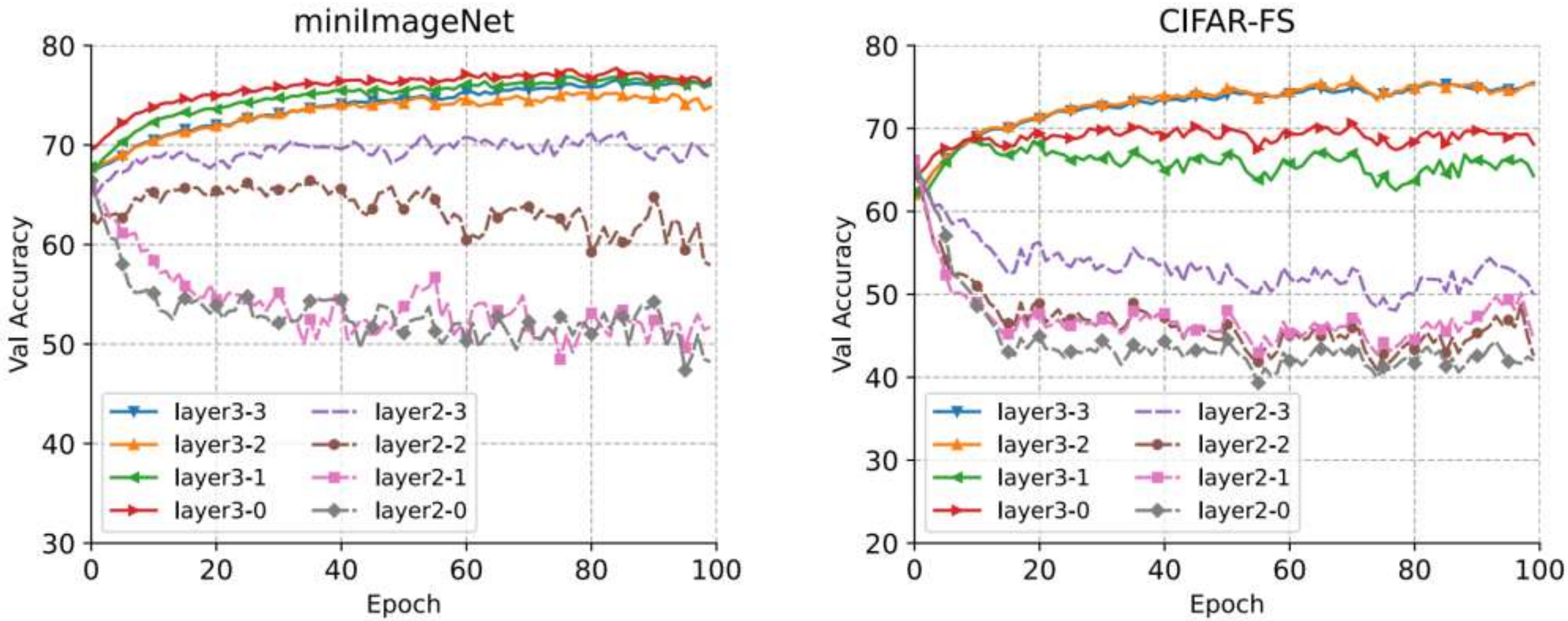



Figure 3. Accuracy vs. different layers to inset prompts. We report 5-way 1-shot accuracy (%) on the validation set of miniImageNet and CIFAF-FS along the meta-training process. The feature extractor has three stages and multiple Transformer layers in each stage.

# Experimental results

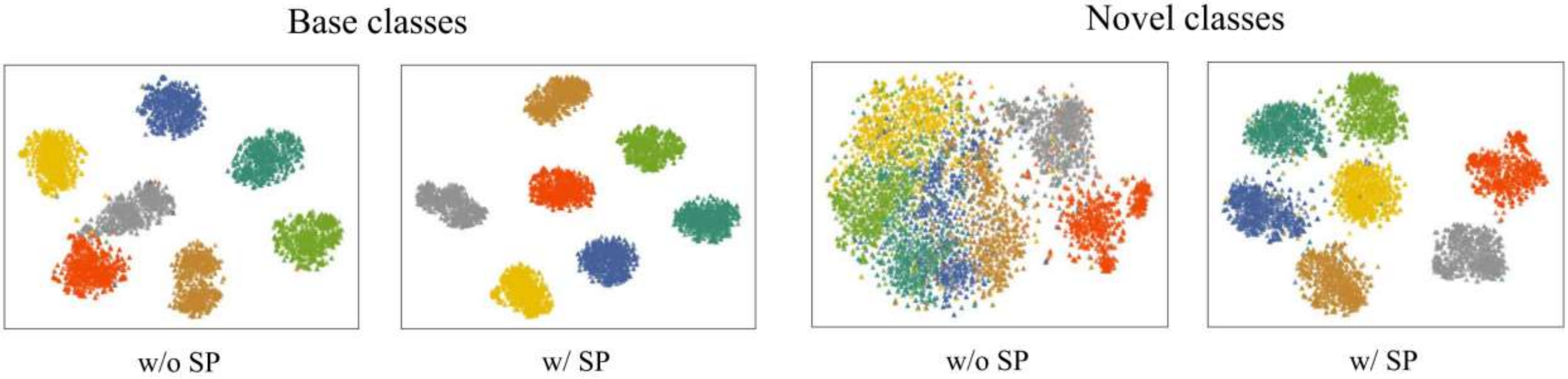



Figure 5. t-SNE results of feature distributions.

# Summary

- We investigate how to use text data to improve the visual feature extraction for few-shot learning.
- We propose a new semantic prompt approach, where text features are used as prompts to adaptively tune the visual features.
- We propose two interaction mechanism, which allow the semantic prompt and visual features to interact along the spatial and the channel dimensions.
- Our approach is evaluated on four datasets with three different text encoders. Experimental results show that using semantic prompt can obtain much more performance gain than previous methods.