# Visual Prompting for Generalized Few-shot Segmentation: A Multi-scale Approach

Mir Rayat Imtiaz Hossain, Mennatullah Siam, Leonid Sigal and James J. Little

*CVPR* (2024)

# Problem

- Prompting for *generalized few-shot* segmentation (GFSS) is **under-explored** (Goal of GFSS: perform well on all base and novel classes)

- Visual prompt tuning (ie: fine-tuning model) is:
  - Easy for **base classes** with **lots of data**, BUT
  - Challenging for **novel classes** with **few examples**

# Objective

Eliminate confusion between base and novel categories when learning new prompts for novel classes.

# Method

Create new prompts for novel classes for *Generalized Few-shot Segmentation* via:

- A multi-scale visual prompting transformer decoder architecture
- A **uni-directional causal attention** mechanism b/n novel and base prompts
- **Transductive prompt-tuning** (on unlabelled test images)

Visual prompts (ie: learnable prompt embeddings) interact with image features at *different scales*
(# visual prompts = # classes)

# Method: Step 1 - Train base visual prompts

1. Initialized randomly

2. Refine through multiple levels of transformer attention

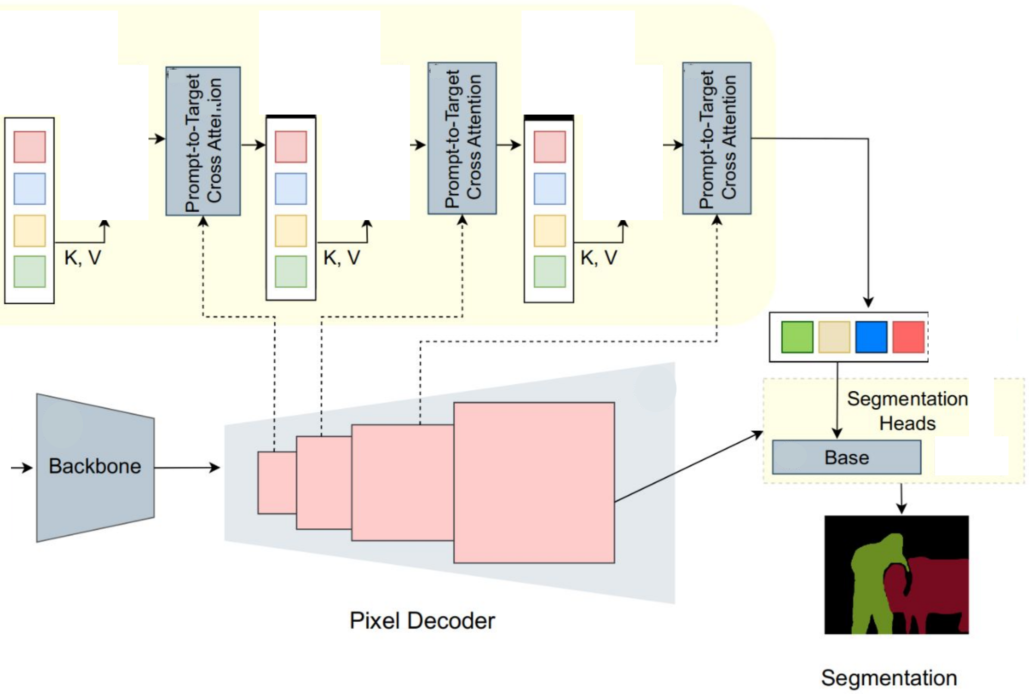
   $$V_B^{(l)} = \mathcal{A}(V_B^{(l-1)}) + \mathcal{C}(V_B^{(l-1)}, F^{(l-1)})$$

3. Define segmentation head

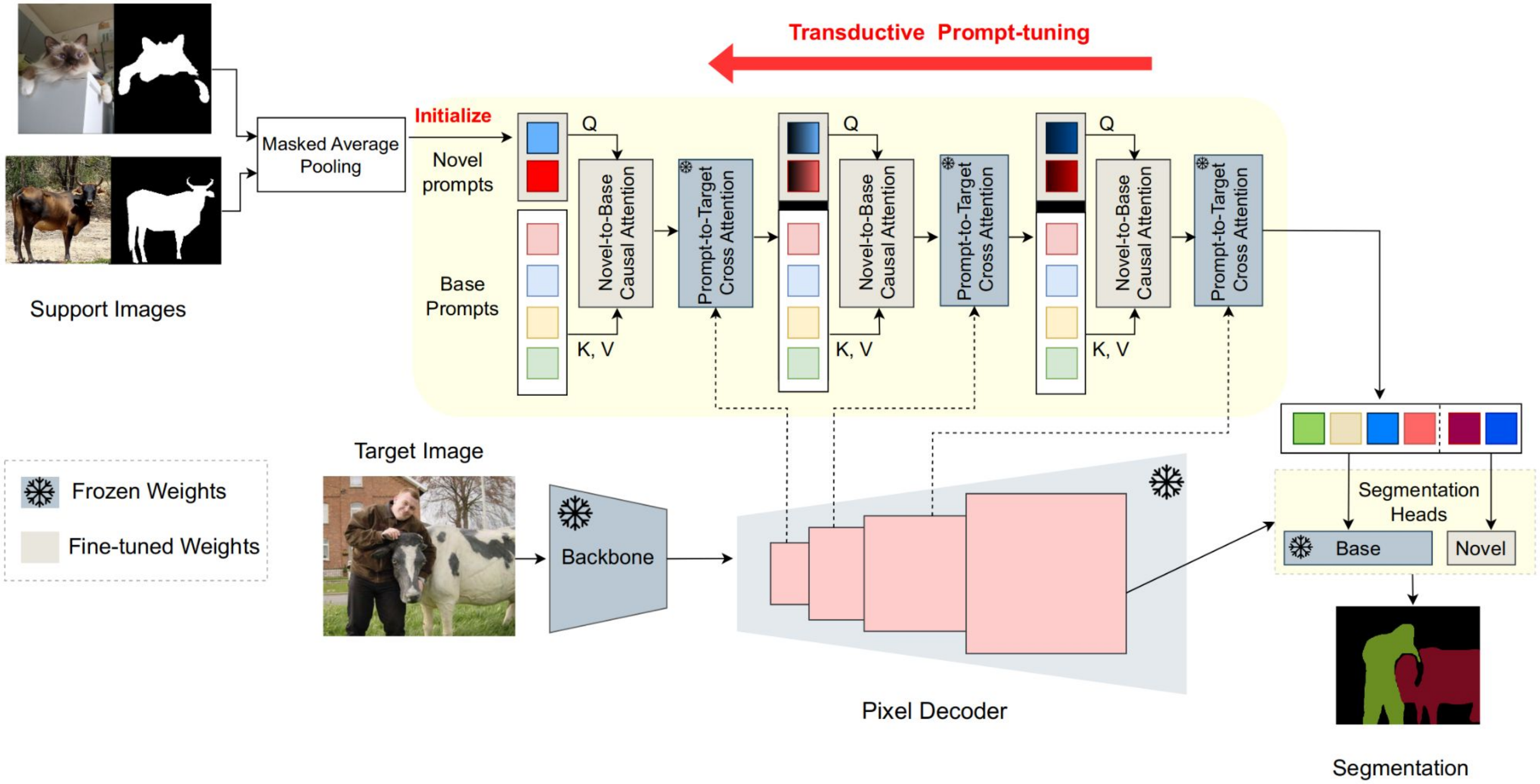
   Computes similarity between pixel features and refined base prompts

# Method: Step 1

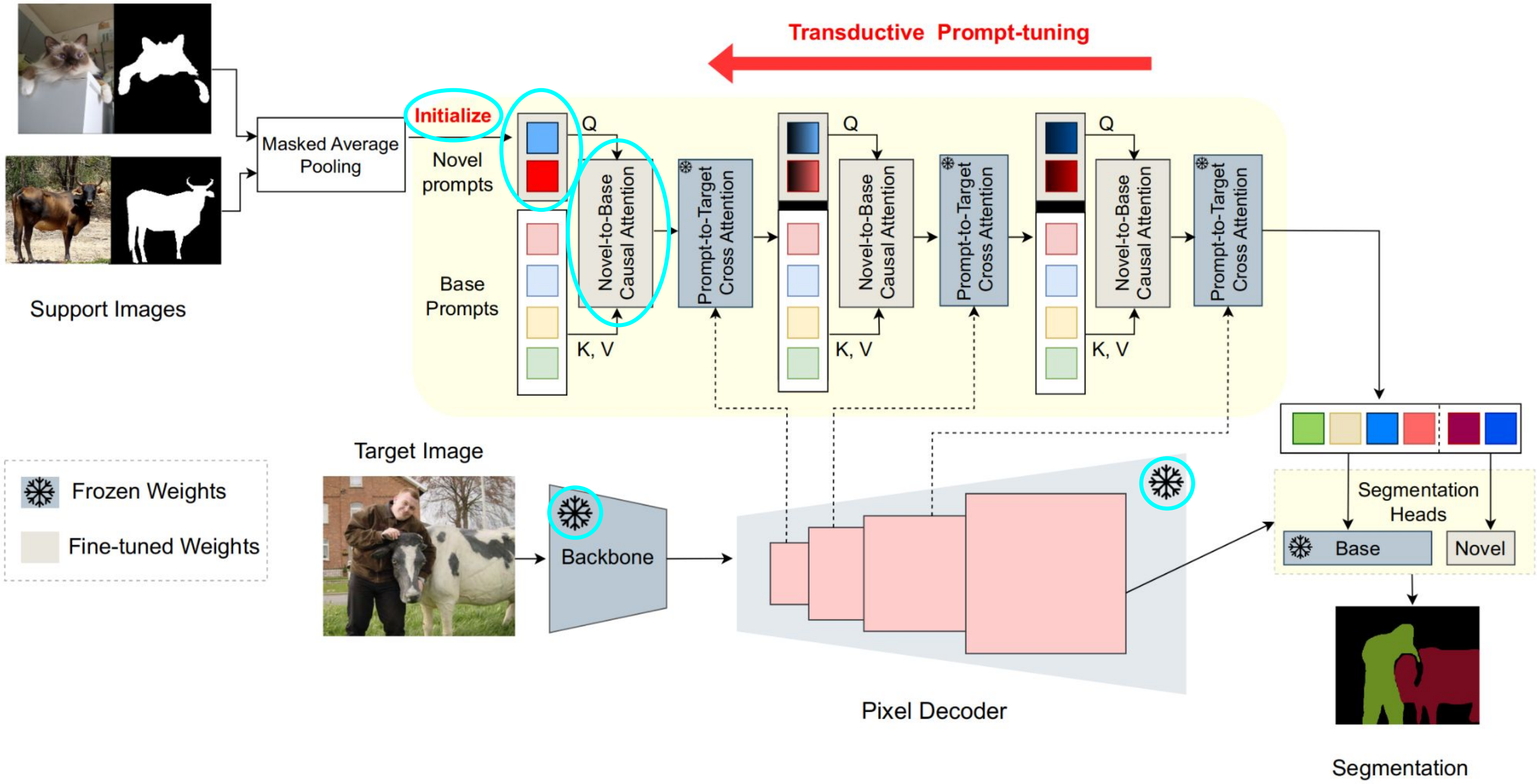
Many examples



# Method: Step 2



# Method: Step 2



# Method: Step 2 - Train novel visual prompts

1. Freeze model except visual prompts
2. Add Novel visual prompts + Novel-to-base causal unidirectional attention module
3. Initialize novel visual prompts w/ average global pooling of masked support images
4. Refine new visual prompts with the added module
5. Add trainable weight to segmentation head

$$
\begin{aligned}
V_N^{(l)} &= \mathcal{CA}(V_B^{(l)}, V_N^{(l)}) \\
V_A^{(l)} &= \left[V_B^{(l)}, V_N^{(l)}\right] \\
V_A^{(l)} &= \mathcal{A}(V_A^{(l-1)}) + \mathcal{C}(V_A^{(l-1)}, F^{(l-1)})
\end{aligned}
$$

# Method

- Inductive objective function (step 1): CE loss
- Transductive objective function (for step 2):

$$\mathcal{L}_{\text{trans.}} = \underbrace{\alpha \mathrm{H}(O|I) - \mathrm{H}(O)}_{\text{(neg.) Mutual information b/n pixel features and prediction}} + \underbrace{\gamma \mathcal{L}_{\text{KD}}}_{\text{KL divergence between predicted prob. of base class at step 1 \& 2}}$$

(neg.) Mutual information b/n pixel features and prediction

KL divergence between predicted prob. of base class at step 1 & 2

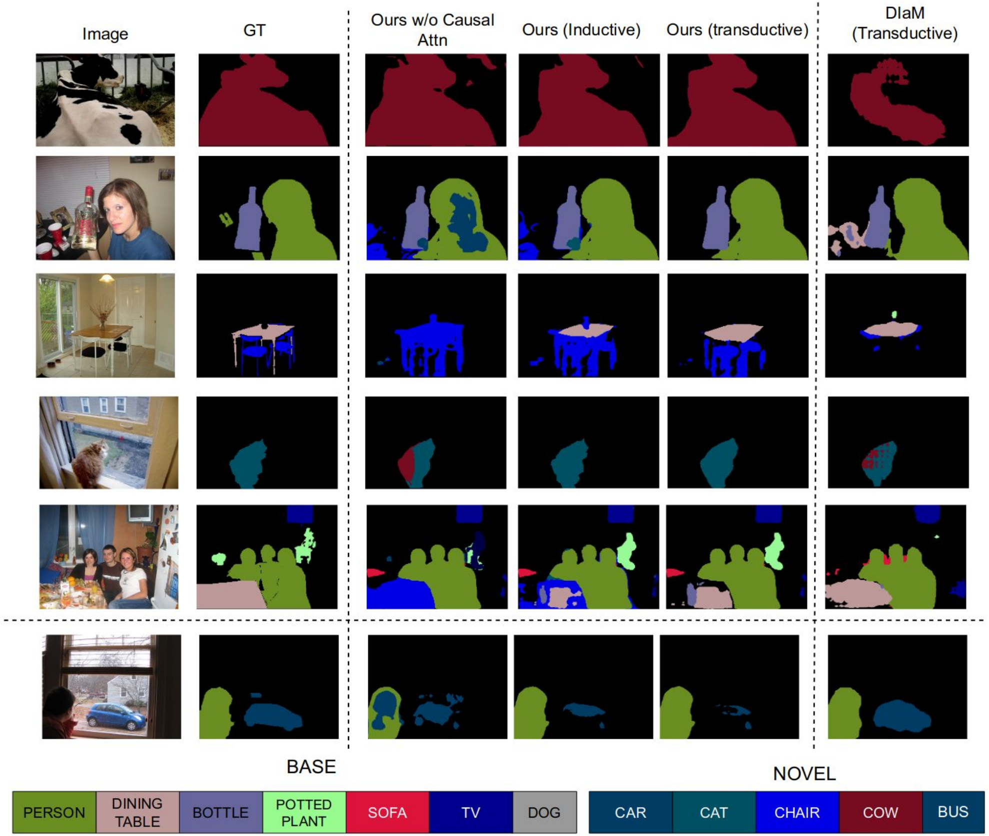
# Experiment

- Datasets: COCO-20i (80 categories) and PASCAL-5i (20 categories)

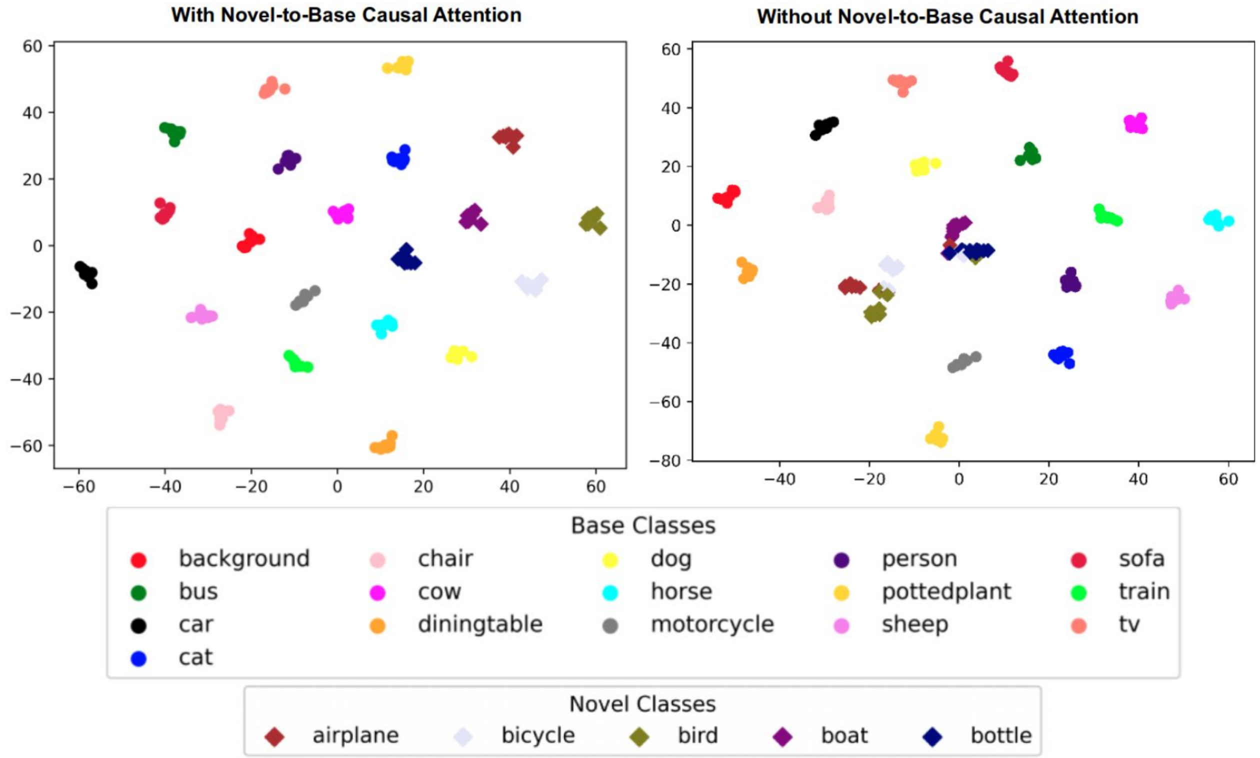
# Quantitative Results

| Method | Learning | PASCAL-$5^i$ 1-shot Base | Novel | Mean | 5-shot Base | Novel | Mean |
|---|---|---|---|---|---|---|---|
| CANeT [50] (CVPR19) | Inductive | 8.73 | 2.42 | 5.58 | 9.05 | 1.52 | 5.29 |
| PFENET [36] (TPAMI20) | Inductive | 8.32 | 2.67 | 5.50 | 8.83 | 1.89 | 5.36 |
| PANET [41] (ICCV19) | Inductive | 31.88 | 11.25 | 21.57 | 32.95 | 15.25 | 24.1 |
| SCL [48] (CVPR21) | Inductive | 8.88 | 2.44 | 5.66 | 9.11 | 1.83 | 5.47 |
| MiB [8] (CVPR20) | Inductive | 63.80 | 8.86 | 36.33 | 68.60 | 28.93 | 48.77 |
| CAPL [37] (CVPR22) | Inductive | 64.80 | 17.46 | 41.13 | 65.43 | 24.43 | 44.93 |
| BAM [19] (CVPR22) | Inductive | 71.60 | 27.49 | 49.55 | 71.60 | 28.96 | 50.28 |
| DIaM (w/o trans.)* [13] (CVPR23) | Inductive | 66.79 | 27.36 | 47.08 | 64.05 | 34.56 | 49.31 |
| POP† [21] (CVPR23) | Inductive | 46.68 | 19.96 | 33.32 | 41.50 | 36.26 | 38.80 |
| **Ours (w/o trans.)** | Inductive | **74.58** | **34.99** | **54.79** | **74.86** | **50.34** | **62.60** |
| RePRI [4] (CVPR21) | Transductive | 20.76 | 10.50 | 15.63 | 34.06 | 20.98 | 27.52 |
| DIaM [13] (CVPR23) | Transductive | 70.89 | 35.11 | 53.00 | 70.85 | 55.31 | 63.08 |
| POP† [21] (CVPR23) | Transductive | 73.92 | 35.51 | 54.72 | 74.78 | 55.87 | 65.33 |
| **Ours (w/ trans.)** | Transductive | **76.39** | **39.83** | **58.11** | **76.42** | **56.12** | **66.27** |

# Qualitative Results on 1-shot segmentation



# Results: Visualization of learned base & novel prompt features



# Results: Ablation on prompt initialization and architecture

W/ causal attention module

| | Causal Attention | Prompt Initialization | Transduction | mIoU | | |
|---|---|---|---|---|---|---|
| | | | | Base | Novel | Mean |
| **(1)** | None | Random | No | 50.18 | 11.22 | 30.70 |
| **(2)** | None | Masked Pooling | No | 50.43 | 11.31 | 30.87 |
| **(3)** | First Layer Only | Masked Pooling | No | 51.53 | 12.26 | 31.90 |
| **(4)** | Separate Weights | Masked Pooling | No | 51.06 | 18.05 | 34.56 |
| **(5)** | Shared Weights | Random | No | 50.75 | 17.41 | 34.08 |
| **(6)** | Shared Weights | Masked Pooling | No | 51.55 | 18.00 | 34.78 |
| **(7)** | Shared Weights | Masked Pooling | Yes | **53.80** | **18.30** | **36.05** |

# Conclusion

- Method that **learns visual prompts** from **few-shot examples** of unseen categories
- Generalizes well to **both seen and unseen** classes

Key components:

- Prompting image features at multiple scales
- Novel-to-base causal attention mechanism that helps contextualize novel class embeddings → reduces confusion with base counterparts
- Applicable to both inductive and transductive settings