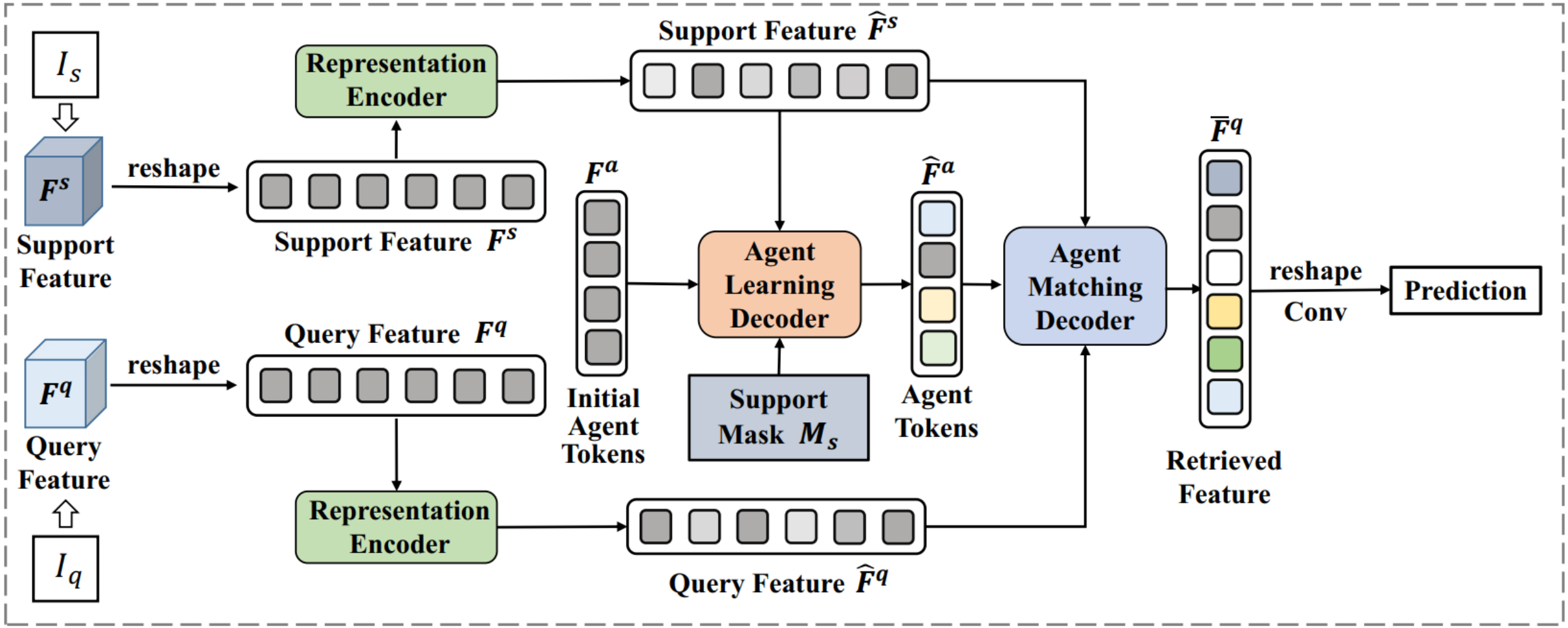
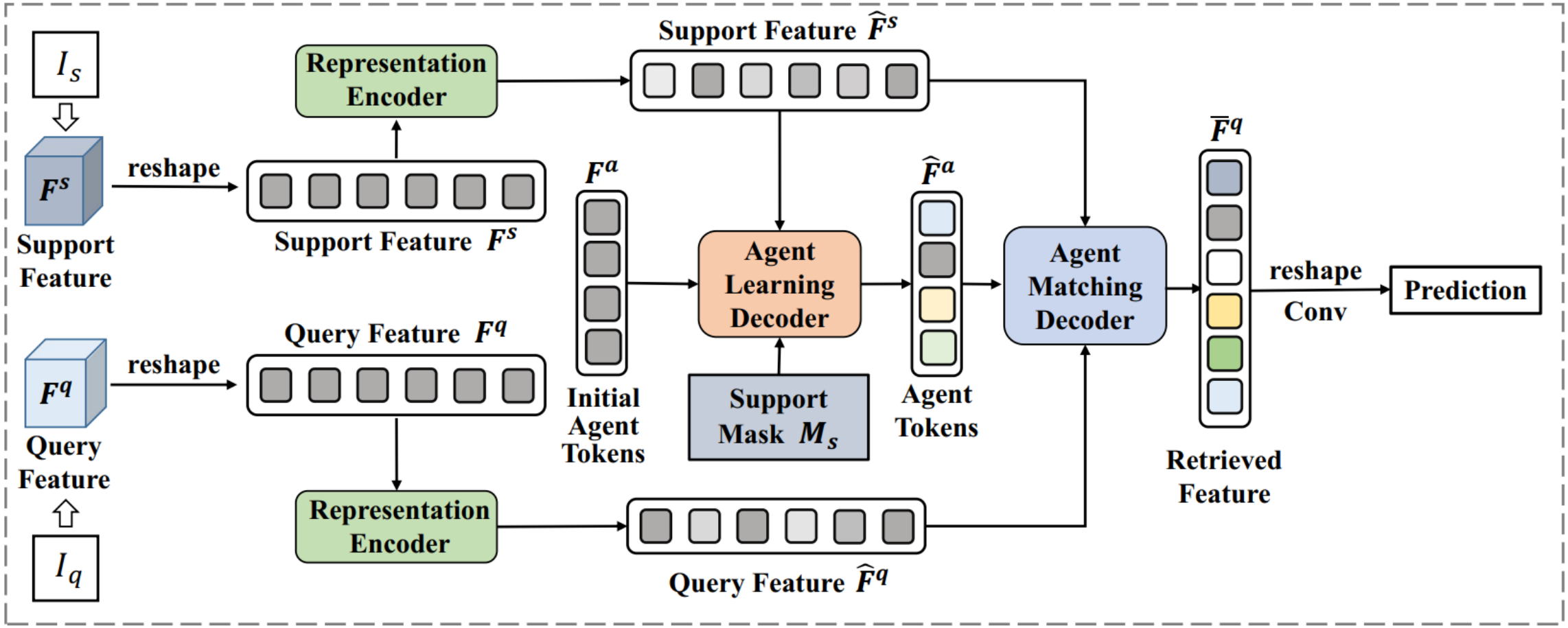
Adaptive Agent Transformer for Few-shot Segmentation



Paper ID - 1397

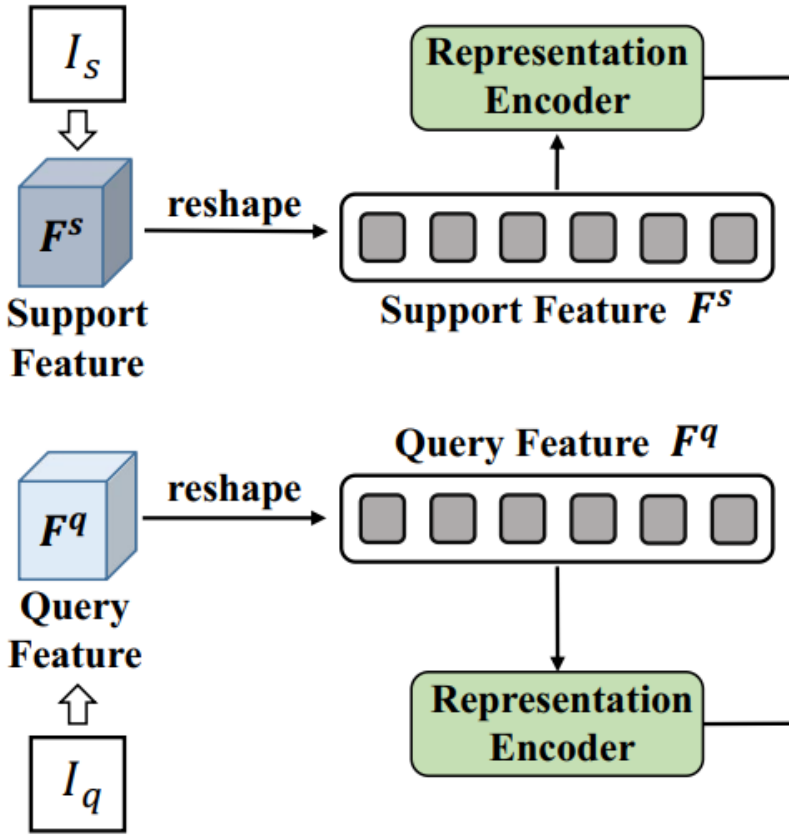
Overview



This work aims to absorb the merits of both prototypical learning and affinity learning formulation via a transformer encoder-decoder architecture, including a **representation encoder**, an **agent learning decoder** and an **agent matching decoder**.

Representation Encoder

We propose the self-attention mechanism to capture the full image context information. Specifically, we aggregate pixel-specific global context to each pixel position to obtain robust context-aware pixel features that can represent object appearance well.



$$\mathbf{Q}^* = \mathbf{F}^* \mathbf{W}_*^{\mathcal{Q}}, \quad \mathbf{K}^* = \mathbf{F}^* \mathbf{W}_*^{\mathcal{K}}, \quad \mathbf{V}^* = \mathbf{F}^* \mathbf{W}_*^{\mathcal{V}}$$

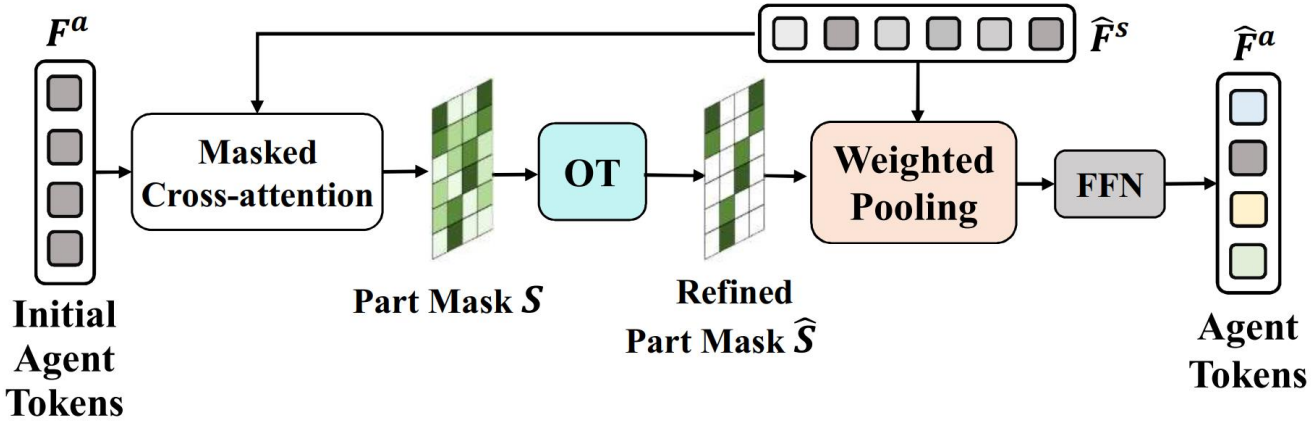
$$\mathbf{W}_*^{\mathcal{Q}} \in \mathbb{R}^{c \times c_k}, \mathbf{W}_*^{\mathcal{K}} \in \mathbb{R}^{c \times c_k}, \mathbf{W}_*^{\mathcal{V}} \in \mathbb{R}^{c \times c_v}$$

$$\hat{\mathbf{F}}^* = \mathrm{Attention}(\mathbf{Q}^*, \mathbf{K}^*, \mathbf{V}^*) = \mathrm{Softmax}(\frac{\mathbf{Q}^* (\mathbf{K}^*)^{\mathsf{T}}}{\sqrt{\mathrm{d_k}}}) \mathbf{V}^*$$

Agent Learning Decoder

We distill support information into condensed agent tokens to establish the bridge between the support and query images. To learn agent tokens well without any explicit supervision, and to make agent tokens capable of dividing different objects into diverse parts in an adaptive manner, we further customize the agent learning decoder according to the three characteristics of context awareness, spatial awareness and diversity.

➢ Context Awareness & Spatial Awareness



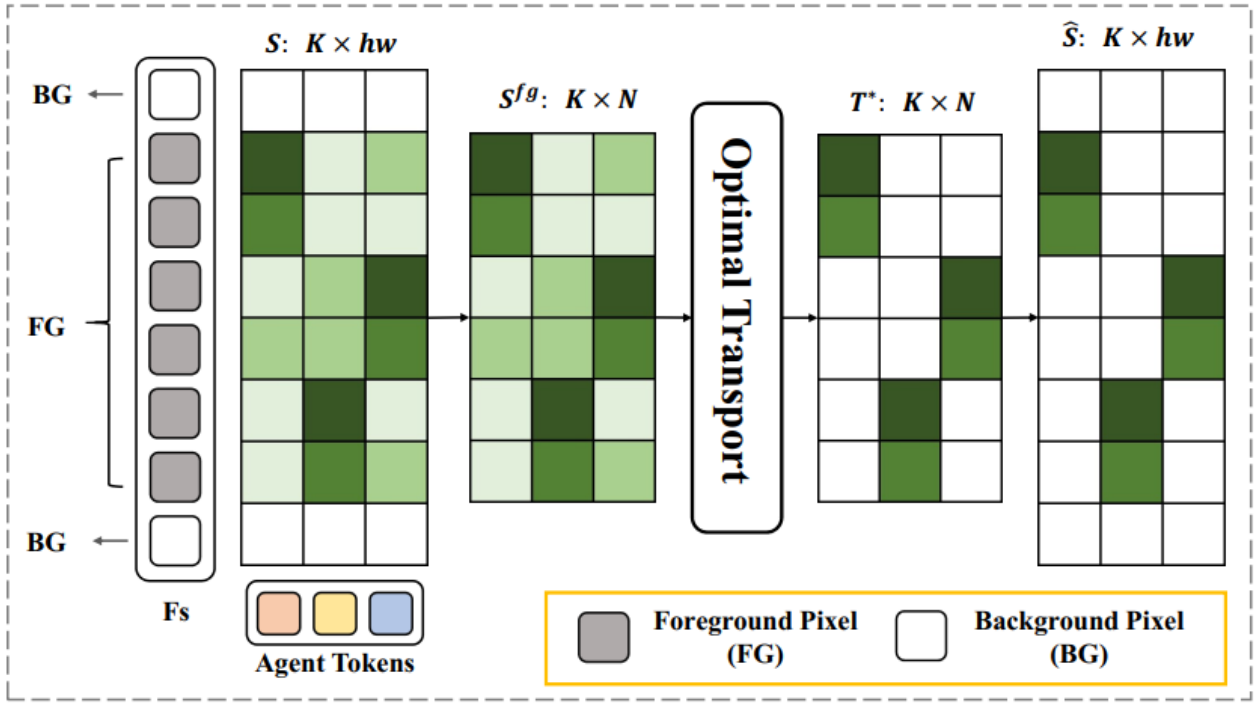
$$\mathbf{S} = \mathrm{Softmax}(\frac{\mathbf{Q}^{\mathrm{a}}(\mathbf{K}^{\mathrm{s}})^{\mathsf{T}}}{\sqrt{\mathrm{d_k}}} + \boldsymbol{\mathcal{M}}), \quad \mathbf{Q}^{\mathrm{a}} = \mathbf{F}^{\mathrm{a}}\mathbf{W}_{\mathrm{a}}^{\mathcal{Q}}, \quad \mathbf{K}^{\mathrm{s}} = \mathbf{F}^{\mathrm{s}}\mathbf{W}_{\mathrm{s}}^{\mathcal{K}}$$

$$\boldsymbol{\mathcal{M}} = \begin{cases} 0, & \text{if} \quad \mathbf{N}(m,n) = 1 \\ -\infty, & \text{otherwise} \end{cases}$$

Agent Learning Decoder

➢ Diversity

We impose the equal partition constraint to expand the discrepancy among part masks. In this case, agent tokens can decompose different target objects into diverse and complementary parts in an adaptive manner.



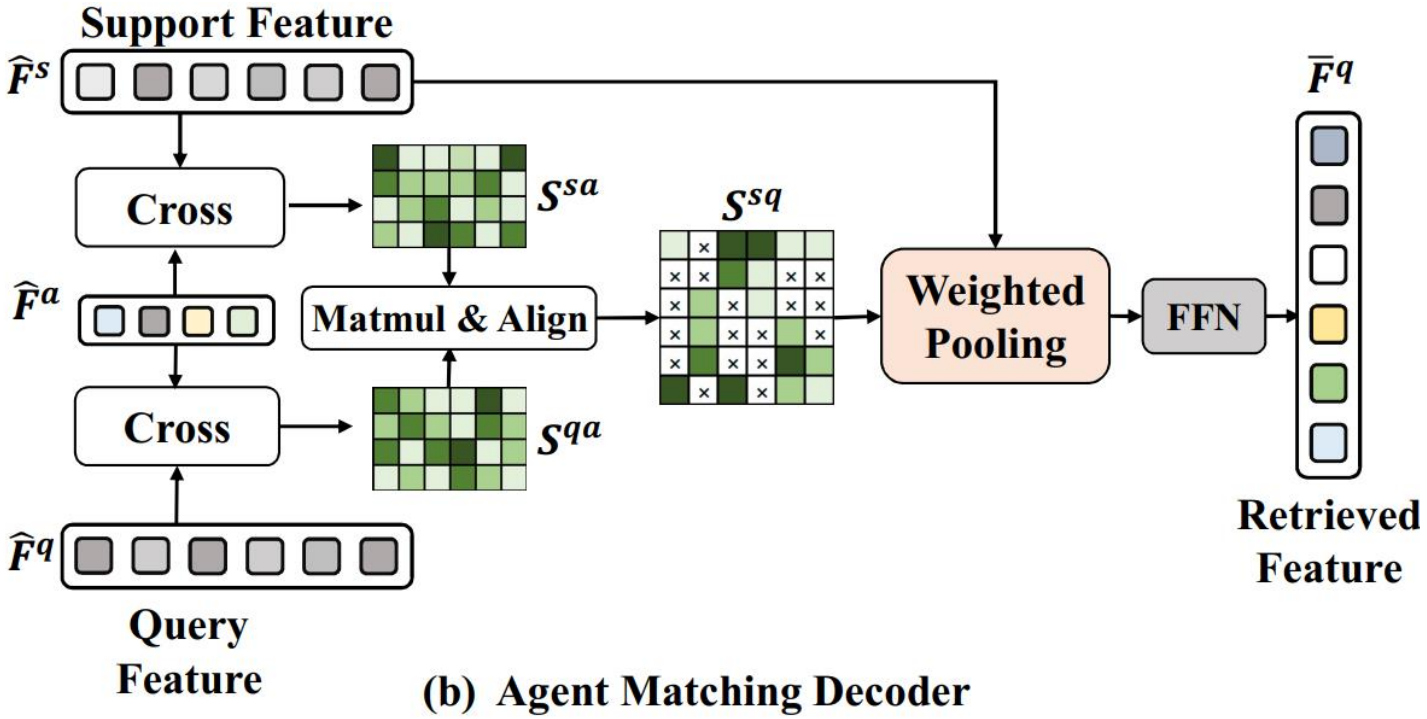
$$\max_{\mathbf{T}\in\mathcal{T}} \operatorname{Tr}\left(\mathbf{T}^{\mathsf{T}}(1-\mathbf{S}^{fg})\right) + \epsilon H(\mathbf{T})$$

$$H(\mathbf{T}) = -\sum_{ij} \mathbf{T}_{ij} \log \mathbf{T}_{ij}$$

$$\mathcal{T} = \left\{ \mathbf{T} \in \mathbb{R}_{+}^{K\times N} \mid \mathbf{T}\mathbb{1} = \frac{1}{K}\cdot\mathbb{1}, \mathbf{T}^{\mathsf{T}}\mathbb{1} = \frac{1}{N}\cdot\mathbb{1} \right\}$$

Agent Matching Decoder

We decompose the massive pixel-level support-query matching matrix into two more manageable matrices based on obtained agent tokens at a light computational cost, and introduce the alignment matrix for filtering out ambiguous matching caused by noisy pixels.



(b) Agent Matching Decoder

$$\mathbf{S}^{as} = \frac{\mathbf{Q}^a(\mathbf{K}^s)^\mathsf{T}}{\sqrt{d_k}}, \quad \mathbf{Q}^a = \hat{\mathbf{F}}^a\mathbf{W}_a^{\mathcal{Q}}, \quad \mathbf{K}^s = \hat{\mathbf{F}}^s\mathbf{W}_s^{\mathcal{K}},$$

$$\mathbf{S}^{qa} = \frac{\mathbf{Q}^q(\mathbf{K}^a)^\mathsf{T}}{\sqrt{d_k}}, \quad \mathbf{Q}^q = \hat{\mathbf{F}}^q\mathbf{W}_q^{\mathcal{Q}}, \quad \mathbf{K}^a = \hat{\mathbf{F}}^a\mathbf{W}_a^{\mathcal{K}},$$

$$\mathbf{S}^{qs} = \mathrm{Softmax}(\mathbf{S}^{\mathrm{sa}}\mathbf{S}^{\mathrm{aq}} + \mathcal{A}),$$

$$\mathcal{A}(i,j) = \begin{cases} 0, & \text{if } \mathrm{argmax_i}\,\mathbf{S}^{\mathrm{as}}(\mathrm{t},\mathrm{i}) = \mathrm{argmax_j}\,\mathbf{S}^{\mathrm{qa}}(\mathrm{j},\mathrm{t}) \\ -\infty, & \text{otherwise} \end{cases}$$

Experiments

➢ Performance on Pascal VOC 2012 and COCO Benchmark

Table 1: Comparison with other state-of-the-art methods for 1-shot and 5-shot segmentation on Pascal-5^i. The mIoU of each fold and the FB-IoU of four folds are reported. Best results in bold.

Method	Backbone	mIoU(1-shot)					FB-IoU	mIoU(5-shot)					FB-IoU
		5^0	5^1	5^2	5^3	Mean	(1-shot)	5^0	5^1	5^2	5^3	Mean	(5-shot)
PANet[ICCV2019] [35]	Vgg-16	42.3	58.0	51.1	41.2	48.1	66.5	51.8	64.6	59.8	46.5	55.7	70.7
FWB[ICCV2019] [26]		47.0	59.6	52.6	48.3	51.9	-	50.9	62.9	56.5	50.1	55.1	-
SG-One[TCYB2020] [46]		40.2	58.4	48.4	38.4	46.3	63.1	41.9	58.6	48.6	39.4	47.1	65.9
PMM[ECCV2020] [40]		47.1	65.8	50.6	48.5	53.0	-	50.0	66.5	51.9	47.6	54.0	-
ASR[CVPR2021] [21]		50.2	66.4	54.3	51.8	55.7	-	53.7	68.5	55.0	54.8	58.0	-
CANet[CVPR2019] [43]	Res-50	52.5	65.9	51.3	51.9	55.4	66.2	55.5	67.8	51.9	53.2	57.1	69.6
PGNet[ICCV2019] [42]		56.0	66.9	50.6	50.4	56.0	69.9	57.7	68.7	52.9	54.6	58.5	70.5
PPNet[ECCV2020] [22]		47.8	58.8	53.8	45.6	51.5	-	58.4	67.8	**64.9**	56.7	62.0	-
PMM[ECCV2020] [40]		55.2	66.9	52.6	50.7	56.3	-	56.3	67.3	54.5	51.0	57.3	-
PFENet[TPAMI2020] [31]		61.7	69.5	55.4	56.3	60.8	73.3	63.1	70.7	55.8	57.9	61.9	73.9
SCLNet[CVPR2021] [41]		63.0	70.0	56.5	57.7	61.8	71.9	64.5	70.9	57.3	58.7	62.9	72.8
ASGNet[CVPR2021] [19]		58.8	67.9	56.8	53.7	59.3	69.2	63.7	70.6	64.2	57.4	63.9	74.2
MMNet[ICCV2021] [36]		62.7	70.2	57.3	57.0	61.8	-	62.2	71.5	57.5	**62.4**	63.4	-
RePRI[CVPR2021] [1]		60.2	67.0	61.7	47.5	59.1	-	64.5	70.8	71.7	60.3	66.8	-
CWT[ICCV2021] [24]		56.3	62.0	59.9	47.2	56.4	-	61.3	68.5	68.5	56.6	63.7	-
SAGNN[CVPR2021] [38]		64.7	69.6	57.0	57.2	62.1	73.2	64.9	70.0	57.0	59.3	62.8	73.3
ASR[CVPR2021] [21]		55.2	70.4	53.4	53.7	58.2	72.9	59.4	71.9	56.9	55.7	61.0	74.1
CMN[ICCV2021] [39]		64.3	70.0	57.4	59.4	62.8	72.3	65.8	70.4	57.6	60.8	63.7	72.8
CyCTR[NIPS2021] [44]		67.8	72.8	58.0	58.0	64.2	-	71.1	73.2	60.5	57.5	65.6	-
AAFormer (Ours)	Res-50	**69.1**	**73.3**	**59.1**	**59.2**	**65.2**	**73.8**	**72.5**	**74.7**	62.0	61.3	**67.6**	**76.2**
FWB[ICCV2019] [26]	Res-101	51.3	64.5	56.7	52.2	56.2	-	54.9	67.4	62.2	55.3	59.9	-
DAN[ECCV2020] [34]		54.7	68.6	57.8	51.6	58.2	62.3	57.9	69.0	60.1	54.9	60.5	63.9
PFENet[TPAMI2020] [31]		60.5	69.4	54.4	55.9	60.1	72.9	62.8	70.4	54.9	57.6	61.4	73.5
ASGNet[CVPR2021] [19]		59.8	67.4	55.6	54.4	59.3	71.7	64.6	71.3	64.2	57.3	64.4	75.2
RePRI[CVPR2021] [1]		59.6	68.6	**62.2**	47.2	59.4	-	66.2	71.4	67.0	57.7	65.6	-
CWT[ICCV2021] [24]		56.9	65.2	61.2	48.8	58.0	-	62.6	70.2	**68.8**	57.2	64.7	-
CyCTR[NIPS2021] [44]		69.3	72.7	56.5	58.6	64.3	72.9	73.5	74.0	58.6	60.2	66.6	75.0
AAFormer (Ours)	Res-101	**69.9**	**73.6**	57.9	**59.7**	**65.3**	**74.9**	**75.0**	**75.1**	59.0	**63.2**	**68.1**	**77.3**

Experiments

➢ Performance on Pascal VOC 2012 and COCO Benchmark

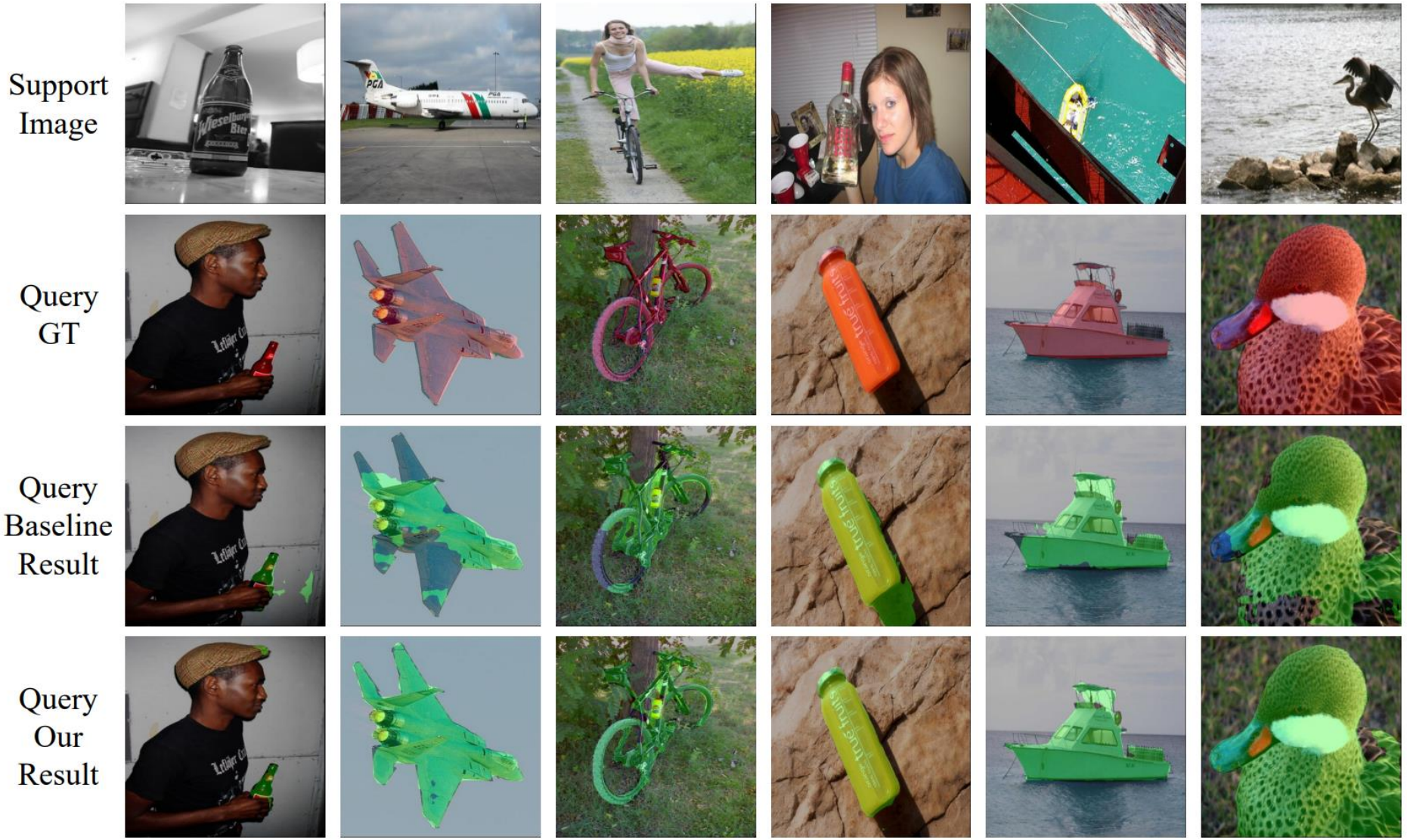
Table 2: Comparison with other state-of-the-art methods for 1-shot and 5-shot segmentation on COCO-20^i. The mIoU of each fold and the FB-IoU of four folds are reported. Best results in bold.

Method	Backbone	mIoU(1-shot)					FB-IoU	mIoU(5-shot)					FB-IoU
		20^0	20^1	20^2	20^3	Mean	(1-shot)	20^0	20^1	20^2	20^3	Mean	(5-shot)
PPNet[ECCV2020] [22]	Res-50	28.1	30.8	29.5	27.7	29.0	-	39.0	40.8	37.1	37.3	38.5	-
PMM[ECCV2020] [40]		29.5	36.8	28.9	27.0	30.6	-	33.8	42.0	33.0	33.3	35.5	-
MMNet[ICCV2021] [36]		34.9	41.0	37.2	37.0	37.5	-	37.0	40.3	39.3	36.0	38.2	-
RePRI[CVPR2021] [1]		31.2	38.1	33.3	33.0	34.0	-	38.5	46.2	40.0	43.6	42.1	-
ASR[CVPR2021] [21]		30.6	36.7	32.7	35.4	33.9	-	33.1	39.5	34.2	36.2	35.8	-
CMN[ICCV2021] [39]		37.9	**44.8**	38.7	35.6	39.3	61.7	42.0	**50.5**	41.0	38.9	43.1	63.3
CyCTR[NIPS2021] [44]		38.9	43.0	39.6	39.8	40.3	-	41.1	48.9	45.2	47.0	45.6	-
FWB[ICCV2019] [26]	Res-101	17.0	18.0	21.0	28.9	21.2	-	19.1	21.5	23.9	30.1	23.7	-
PFENet[TPAMI2020] [31]		34.3	33.0	32.3	30.1	32.4	58.6	38.5	38.6	38.2	34.3	37.4	61.9
SCLNet[CVPR2021] [41]		36.4	38.6	37.5	35.4	37.0	-	38.9	40.5	41.5	38.7	39.9	-
CWT[ICCV2021] [24]		30.3	36.6	30.5	32.2	32.4	-	38.5	46.7	39.4	43.2	42.0	-
SAGNN[CVPR2021] [38]		36.1	41.0	38.2	33.5	37.2	60.9	40.9	48.3	42.6	38.9	42.7	63.4
AAFormer (Ours)	Res-50	**39.8**	44.6	**40.6**	**41.4**	**41.6**	**67.7**	**42.9**	50.1	**45.5**	**49.2**	**46.9**	**68.2**

- We consistently observe that our AAFormer outperforms all previous models under both 1-shot and 5-shot settings, which strongly proves the effectiveness of our method.
- Compared with some recent prototypical learning methods (e.g., ASGNet), our method achieves a large margin of 6.0% and 3.7% in mIoU. This is because the agent matching decoder in our method further explores the pixel-wise support information and contributes to accurate segmentation.

Experiments

➢ Qualitative Comparison with Baseline



Contributions

- We propose an Adaptive Agent Transformer (AAFormer) for the few-shot segmentation in a unified framework. Specifically, we design the representation encoder to acquire global context-aware pixel features, the agent learning decoder to condense support information into agent tokens for bridging the support and query images, and the agent matching decoder to decompose the direct pixel-level matching matrix into two more computationally-friendly matrices for suppressing the noisy pixels.
- To the best of our knowledge, this is the first work to absorb the merits of both prototypical learning and affinity learning formulation by modeling adaptive agent tokens for pixel-level matching. To learn agent tokens well without any explicit supervision, and to make agent tokens capable of dividing different objects into diverse parts in an adaptive manner, we further customize the agent learning decoder according to the three characteristics of context awareness, spatial awareness and diversity.
- Extensive experimental results with two different backbones on two challenging benchmarks demonstrate that our AAFormer performs favorably against state-of-the-art FSS methods

Thank You