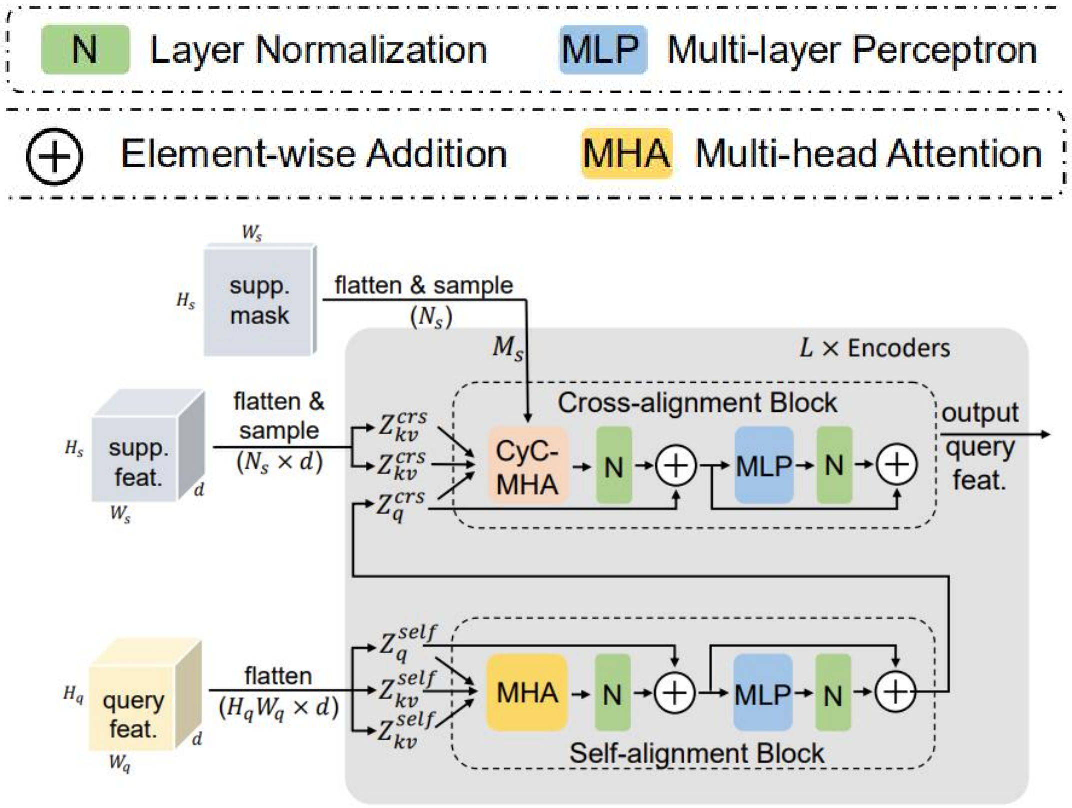
01 Few-Shot Segmentation via Cycle-Consistent Transformer

Few-Shot Segmentation via Cycle-Consistent Transformer

Gengwei Zhang[1,2]*, Guoliang Kang[3], Yi Yang[4], Yunchao Wei[5,6†]
[1] Baidu Research
[2] ReLER, Centre for Artificial Intelligence, University of Technology Sydney
[3] University of Texas, Austin
[4] CCAI, College of Computer Science and Technology, Zhejiang University
[5] Institute of Information Science, Beijing Jiaotong University
[6] Beijing Key Laboratory of Advanced Information Science and Network
{zgwdavid, kgl.prml, wychao1987, yee.i.yang}@gmail.com

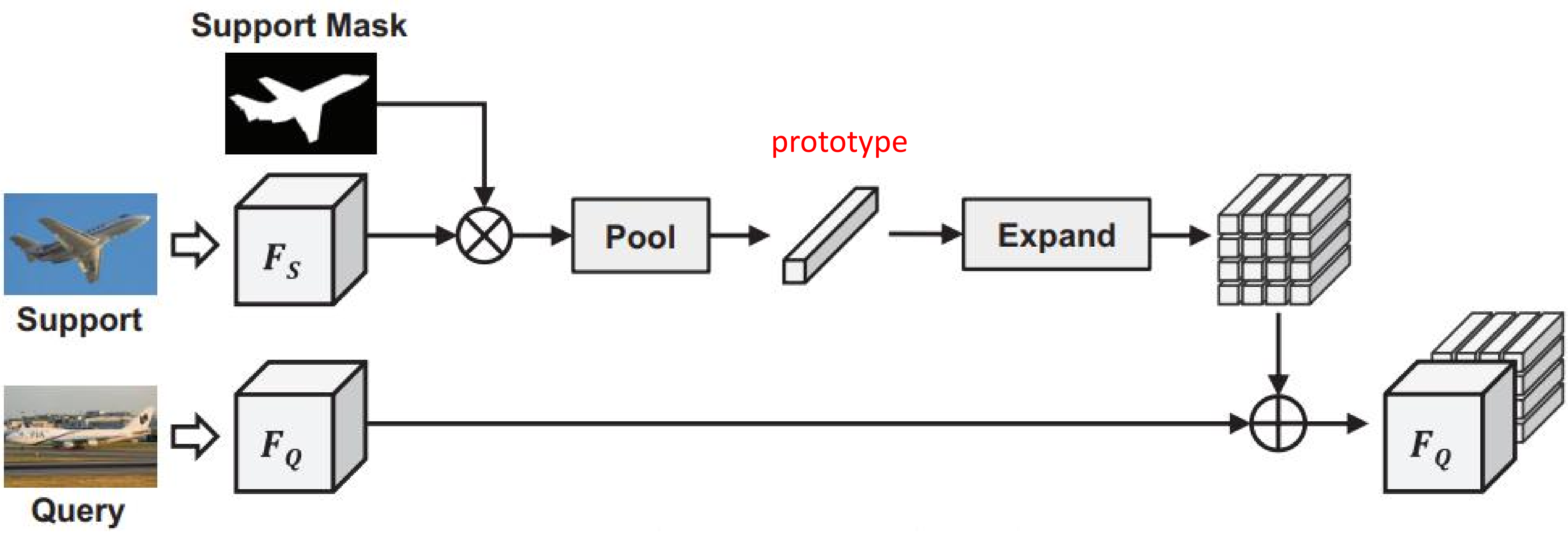


- **Neural Information Processing Systems (NIPS 2021)**

01 Background for Few-Shot Segmentation

- **Motivation**
  - Deep learning based computer vision systems have largely depended on large-scale training sets
  - Deep networks mostly work with predefined classes and are incapable of generalizing to new ones

Few shot: learn how to recognize novel objects after seeing only a handful of exemplars

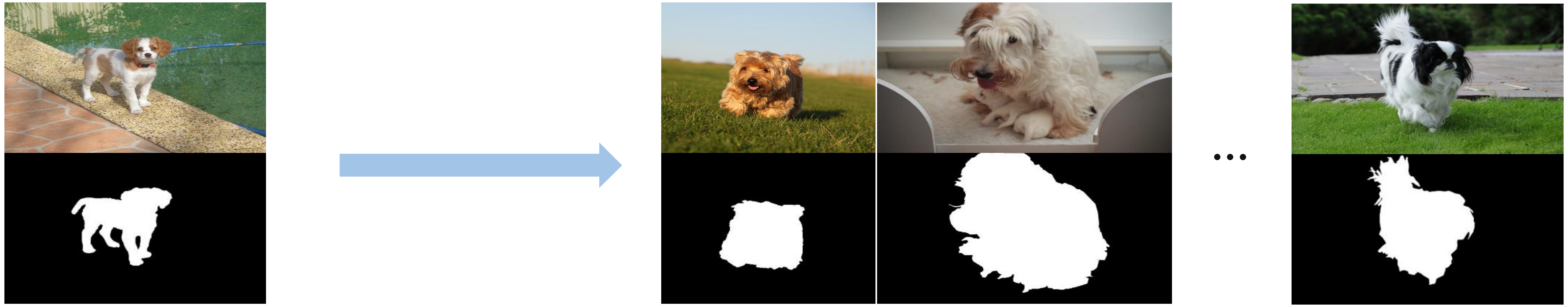


Feature from Res2 + Res3

01 Background for Few-Shot Segmentation

- **Implementation**
  - Divide the dataset into � � � � � and � � � � � � � � , � � � � � ∩ � � � � � � � � = ∅
  - **One-shot segmentation** and ***k*-shot segmentation**

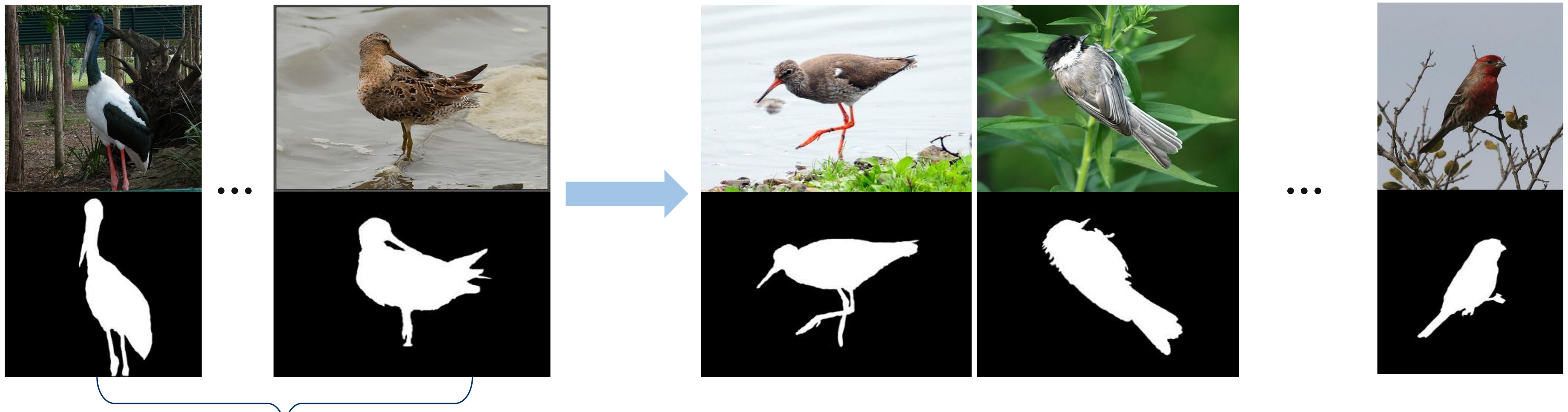
One-shot Segmentation:



Support Image

Query Images

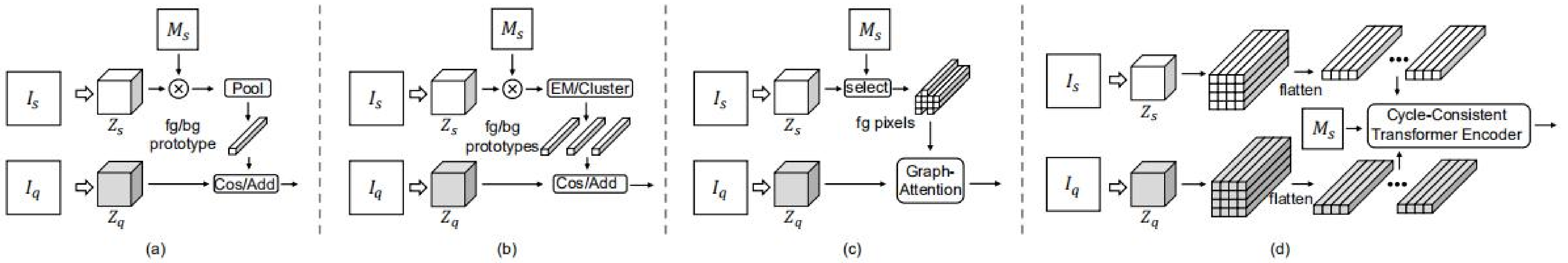
k-shot Segmentation:



k Support Images

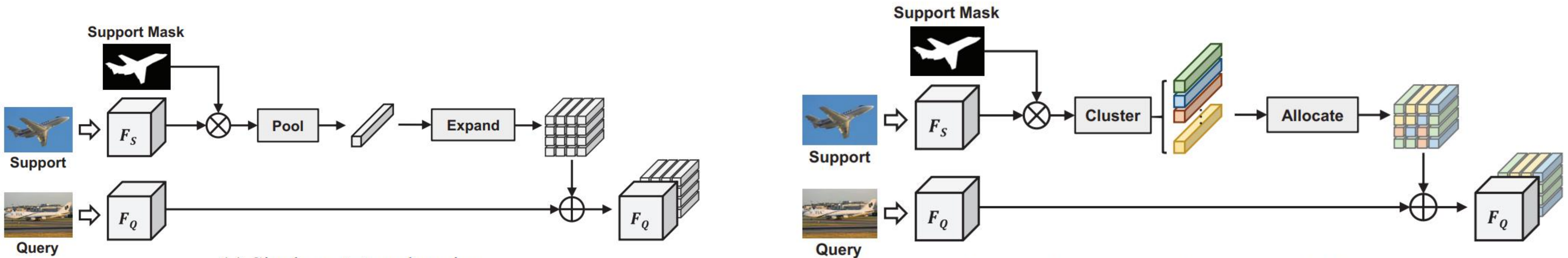
Query Images

01 Few-Shot Segmentation via Cycle-Consistent Transformer

- 

$= \varnothing$

Figure 1: Different learning frameworks for few-shot segmentation, from the perspective of ways to utilize support information. (a) Class-wise mean pooling based method. (b) Clustering based method. (c) Foreground pixel attention method. (d) Our Cycle-Consistent TRansformer (CyCTR) framework that enables all beneficial support pixel-level features (foreground and background) to be considered.



(a) Single prototype learning

(b) Adaptive prototype learning and allocation

Adaptive Prototype Learning and Allocation for Few-Shot Segmentation(CVPR21)

01 Few-Shot Segmentation via Cycle-Consistent Transformer

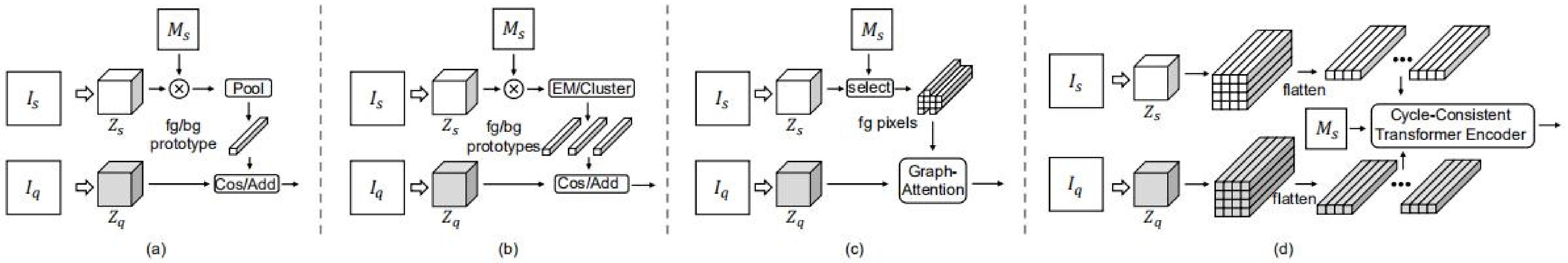
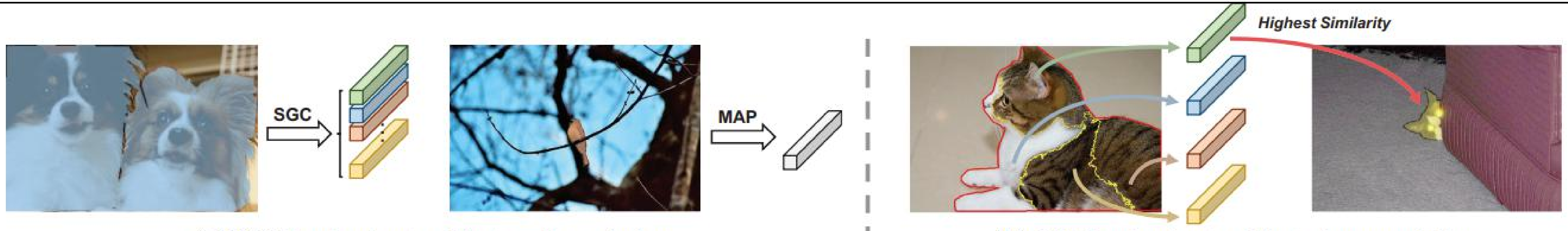
- 

Figure 1: Different learning frameworks for few-shot segmentation, from the perspective of ways to utilize support information. (a) Class-wise mean pooling based method. (b) Clustering based method. (c) Foreground pixel attention method. (d) Our Cycle-Consistent TRansformer (CyCTR) framework that enables all beneficial support pixel-level features (foreground and background) to be considered.

 = ∅



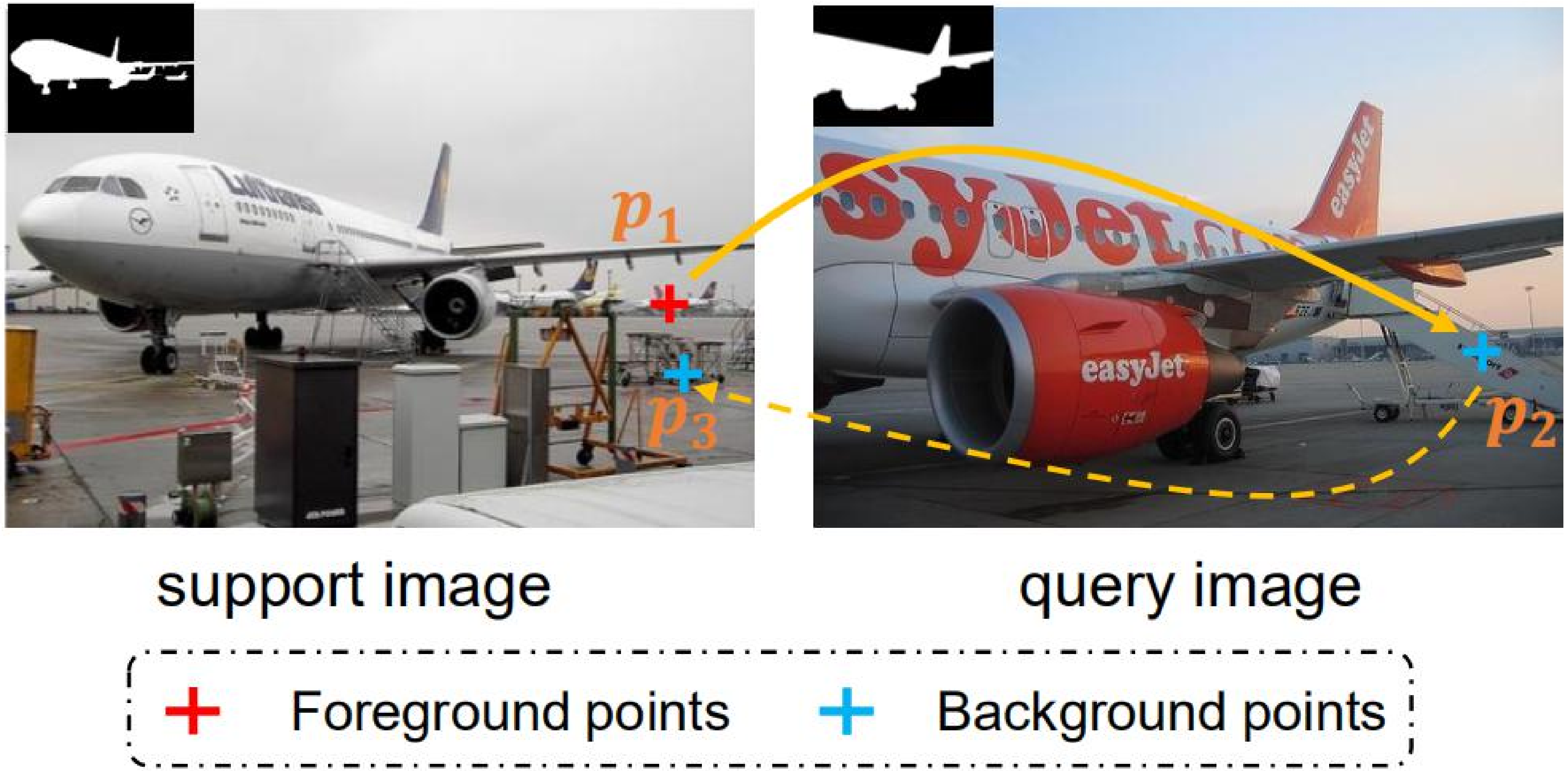
(a) SGC is adaptive to object scale variation

(b) GPA is adaptive to object shape variation

Adaptive Prototype Learning and Allocation for Few-Shot Segmentation(CVPR21)

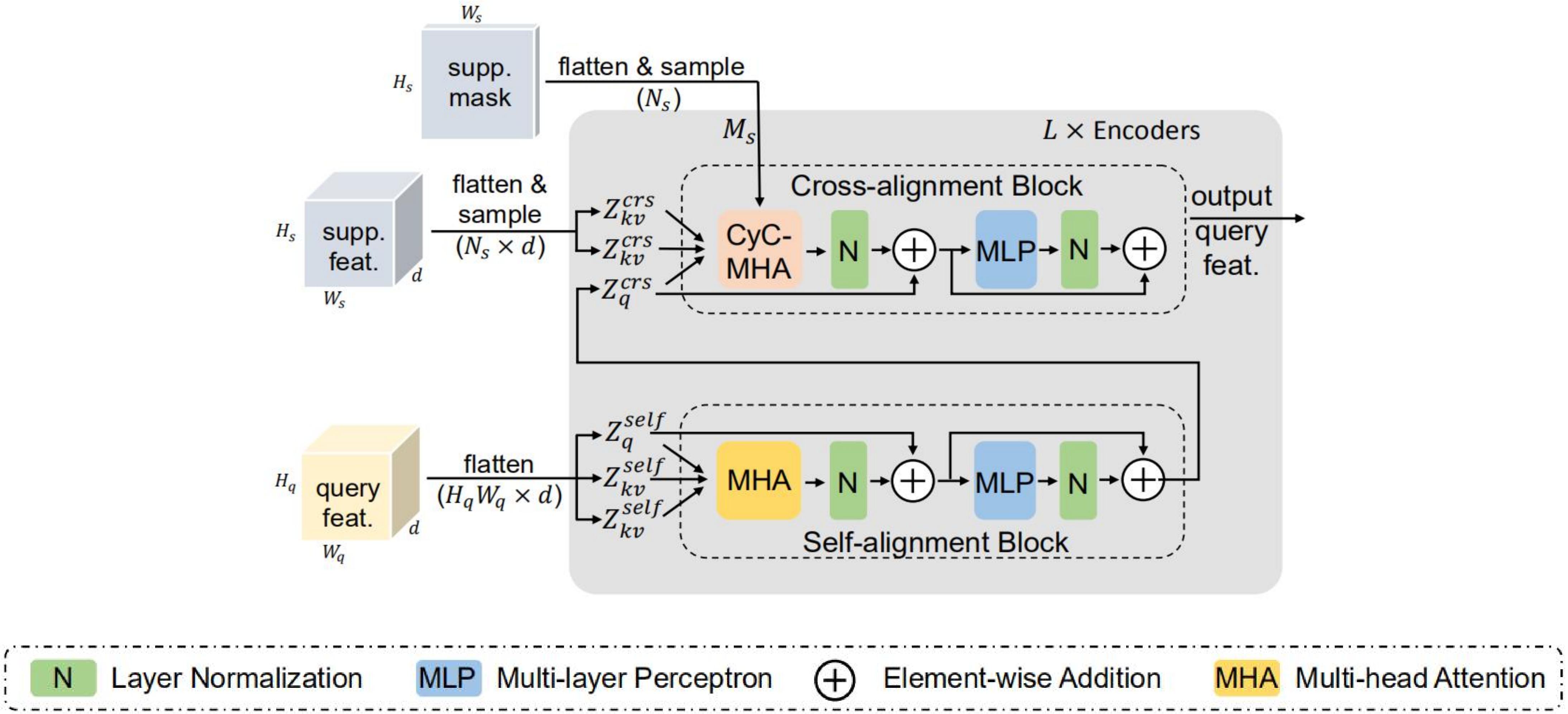
01 Few-Shot Segmentation via Cycle-Consistent Transformer

- **Motivation**
  - Many pixel-level **support features are quite different from the query ones**, and thus may confuse the attention.



01 Few-Shot Segmentation via Cycle-Consistent Transformer

- **Framework of the proposed Cycle-Consistent TRansformer (CyCTR)**
  - **Self-alignment block** for utilizing global context within the query feature map
  - **Cross-alignment block** for aggregate information from support images



01 Few-Shot Segmentation via Cycle-Consistent Transformer

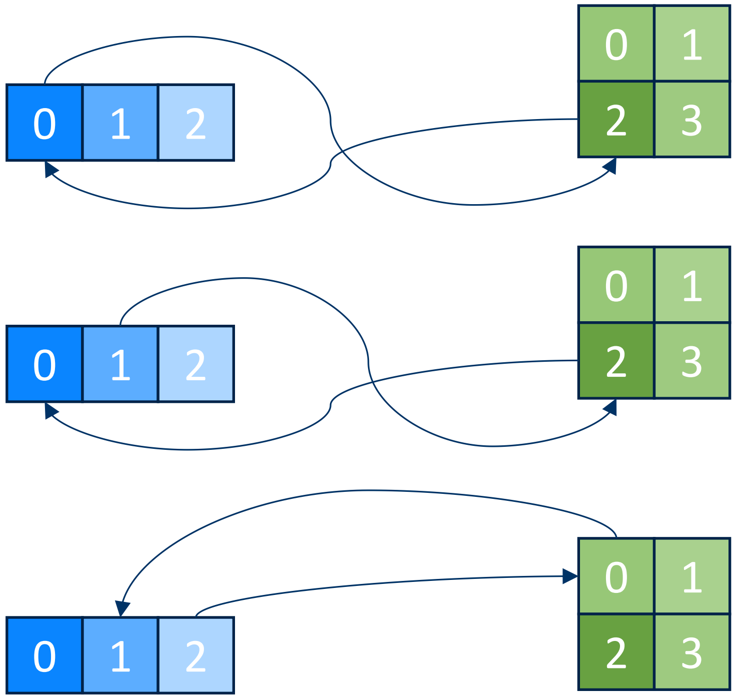
- **Cycle-Consistent Attention（ Cross-alignment block ）**

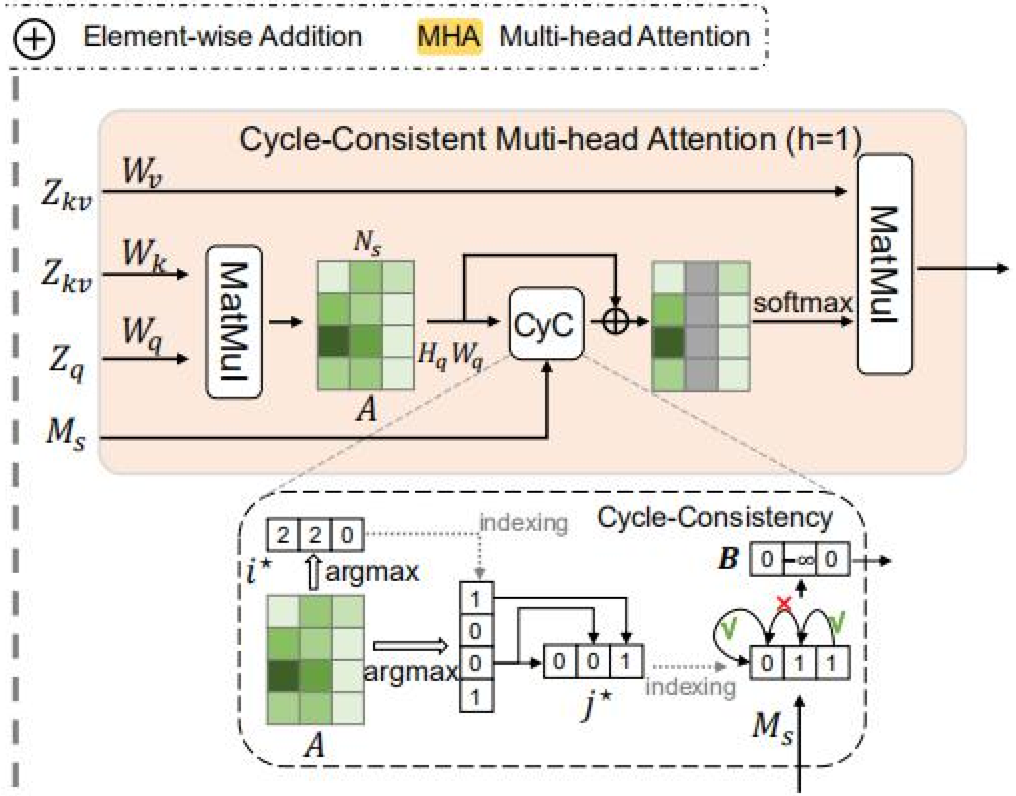
$$A = \frac{QK^T}{\sqrt{d}}, A \in \mathbb{R}^{H_q W_q \times N_s}$$

$$i^\star = \operatorname*{argmax}_i A_{(i,j)}, \qquad j^\star = \operatorname*{argmax}_j A_{(i^\star,j)}$$

$$B_j = \begin{cases} 0, & \text{if} M_{s(j)} = M_{s(j^\star)} \\ -\infty, & \text{if} M_{s(j)} \neq M_{s(j^\star)} \end{cases}$$

$$\mathrm{CyCAtten}(Q_i, K_i, V_i) = \mathrm{softmax}(A_{(i)} + B)V,$$





01 Few-Shot Segmentation via Cycle-Consistent Transformer

- **Mask-guided sparse sampling and K-shot Setting**

$$N_{fg} \leq \frac{N_s}{2} \ , N_s \leq kH_sW_s$$

$$N_{bg} = N_s - N_{fg}$$

With a proper ***Ns***, the sampling operation reduces the computational complexity, and makes our algorithm more scalable with the increase of spatial size of support images.

01 Few-Shot Segmentation via Cycle-Consistent Transformer

- **Self-alignment block**
  - Refering to deformable detr

$$\mathrm{PredAtten}(Q_r, V_r) = \sum_{g}^{P} \mathrm{softmax}(A')_{(r,g)} V_{r+\Delta_{(r,g)}},$$

$$\mathrm{Atten}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d}})V,$$

Compare with original attention

$$\Delta = f(Q + \mathrm{Coord}) \text{ and } A' = g(Q + \mathrm{Coord}), \quad A' \in \mathbb{R}^{H_q W_q \times P}$$

$f(\cdot)$ and $g(\cdot)$ are two fully connected layers that predict the offsets[3] and attention weights.

01 Few-Shot Segmentation Setting

- **Dataset**
  - **Pascal-5^{i}**: 20 categories, 15 classes are used for training and 5 classes for test.
  - **COCO- 20^{i}**: 80 categories, 60 classes are used for training and 20 classes for test.
- **Evaluation Metric**
  - mIoU
  - FB-IoU

01 Few-Shot Segmentation via Cycle-Consistent Transformer

- **Results**

Table 1: Comparison with other state-of-the-art methods for 1-shot and 5-shot segmentation on PASCAL-5^i using the mIoU (%) evaluation metric. Best results are shown in bold.

Method	Backbone	1-shot					5-shot				
		5^0	5^1	5^2	5^3	Mean	5^0	5^1	5^2	5^3	Mean
PANet [35]	Vgg-16	42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7
FWB [23]		47.0	59.6	52.6	48.3	51.9	50.9	62.9	56.5	50.1	55.1
SG-One [45]		40.2	58.4	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1
RPMM [41]		47.1	65.8	50.6	48.5	53.0	50.0	66.5	51.9	47.6	54.0
CANet [44]	Res-50	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1
PGNet [43]		56.0	66.9	50.6	50.4	56.0	57.7	68.7	52.9	54.6	58.5
RPMM [41]		55.2	66.9	52.6	50.7	56.3	56.3	67.3	54.5	51.0	57.3
PPNet [18]		47.8	58.8	53.8	45.6	51.5	58.4	67.8	**64.9**	56.7	62.0
PFENet [30]		61.7	69.5	55.4	56.3	60.8	63.1	70.7	55.8	**57.9**	61.9
CyCTR (Ours)	Res-50	**67.8**	**72.8**	**58.0**	**58.0**	**64.2**	**71.1**	**73.2**	60.5	57.5	**65.6**
FWB [23]	Res-101	51.3	64.5	56.7	52.2	56.2	54.9	67.4	**62.2**	55.3	59.9
DAN [34]		54.7	68.6	**57.8**	51.6	58.2	57.9	69.0	60.1	54.9	60.5
PFENet [30]		60.5	69.4	54.4	55.9	60.1	62.8	70.4	54.9	57.6	61.4
CyCTR (Ours)	Res-101	**69.3**	**72.7**	56.5	**58.6**	**64.3**	**73.5**	**74.0**	58.6	**60.2**	**66.6**

01 Few-Shot Segmentation via Cycle-Consistent Transformer

- **Results**

Table 2: Comparison with other state-of-the-art methods for 1-shot and 5-shot segmentation on COCO-20^i using the mIoU (%) evaluation metric. Best results are shown in bold.

Method	Backbone	1-shot					5-shot				
		20^0	20^1	20^2	20^3	Mean	20^0	20^1	20^2	20^3	Mean
FWB [23]	Res-101	19.9	18.0	21.0	28.9	21.2	19.1	21.5	23.9	30.1	23.7
PPNet [18]	Res-50	28.1	30.8	29.5	27.7	29.0	39.0	40.8	37.1	37.3	38.5
RPMM [41]	Res-50	29.5	36.8	29.0	27.0	30.6	33.8	42.0	33.0	33.3	35.5
PFENet [30]	Res-101	34.3	33.0	32.3	30.1	32.4	38.5	38.6	38.2	34.3	37.4
CyCTR (Ours)	Res-50	**38.9**	**43.0**	**39.6**	**39.8**	**40.3**	**41.1**	**48.9**	**45.2**	**47.0**	**45.6**

01 Few-Shot Segmentation via Cycle-Consistent Transformer

- **Results**

Table 4: Ablation studies that validate the effectiveness of each component in our Cycle-Consistent TRansformer. The first result is obtained by our baseline (see Section 4.2 for details).

self-alignment	cross-alignment	CyCTR (pred)	CyCTR (fg. only)	CyCTR	mIoU (%)
					58.8
✓					61.6
✓	✓				61.2
✓	✓	✓			61.9
✓	✓		✓		62.0
✓	✓			✓	62.8