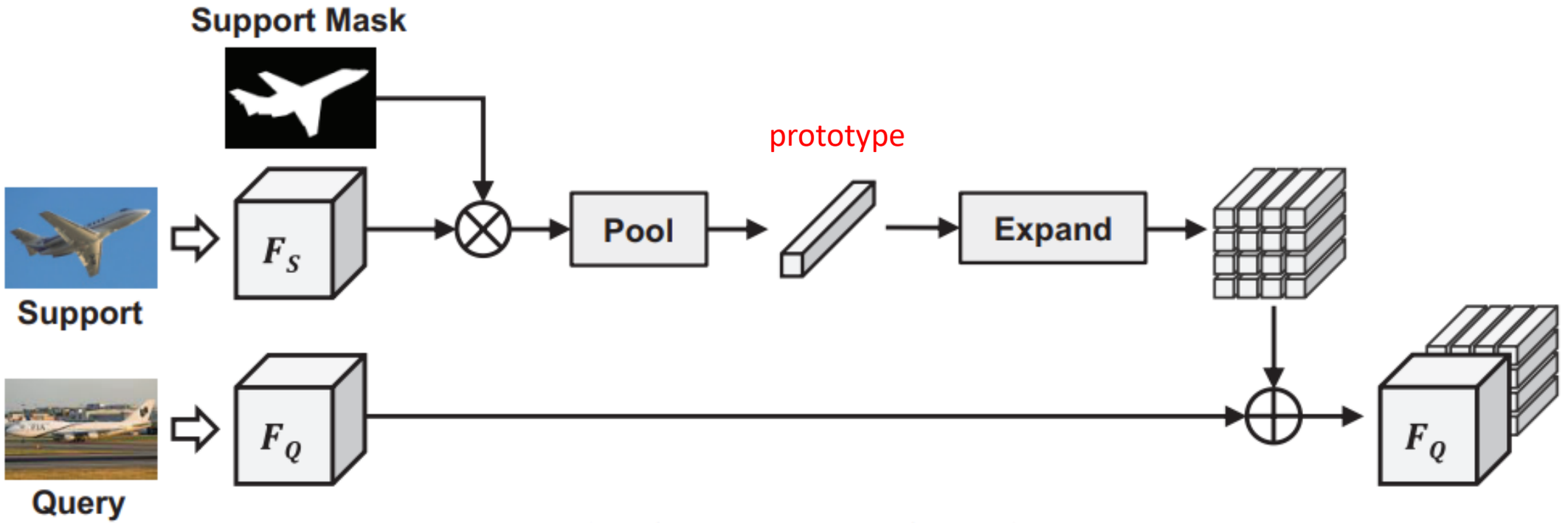
01 Background for Few-Shot Segmentation

- **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

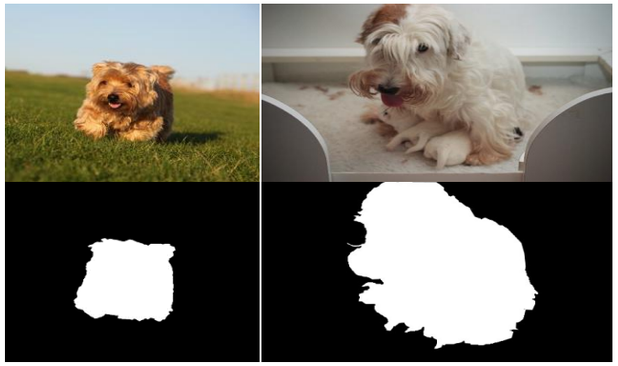


01 Background for Few-Shot Segmentation

- **Implementation**
  - Divide the dataset into $\boldsymbol{Cseen}$ and $\boldsymbol{Cunseen}$ ($\boldsymbol{Cbase}$ and $\boldsymbol{Cnovel}$), $Cseen \cap Cunseen = \emptyset$
  - **One-shot segmentation** and $\boldsymbol{k}$**-shot segmentation**

One-shot Segmentation:

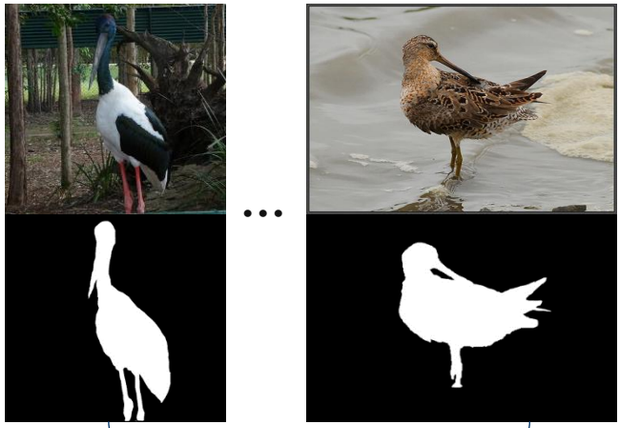
Support Image



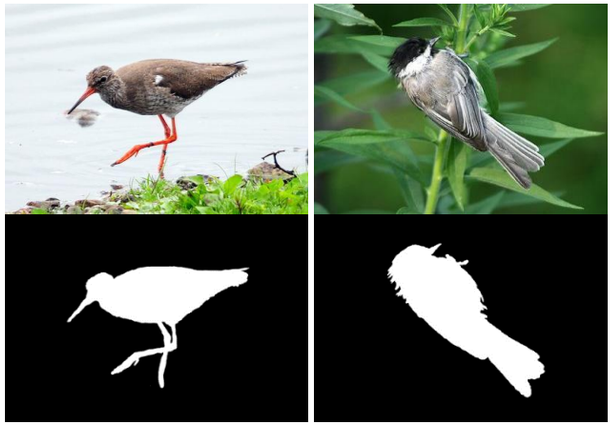
…

Query Images

$\boldsymbol{k}$-shot Segmentation:



$\boldsymbol{k}$ Support Images



…

Query Images

02 Generalized Few-shot Semantic Segmentation

Zhuotao Tian[1] Xin Lai[1] Li Jiang[2] Shu Liu[3] Michelle Shu[4] Hengshuang Zhao[5,6] Jiaya Jia[1,3]

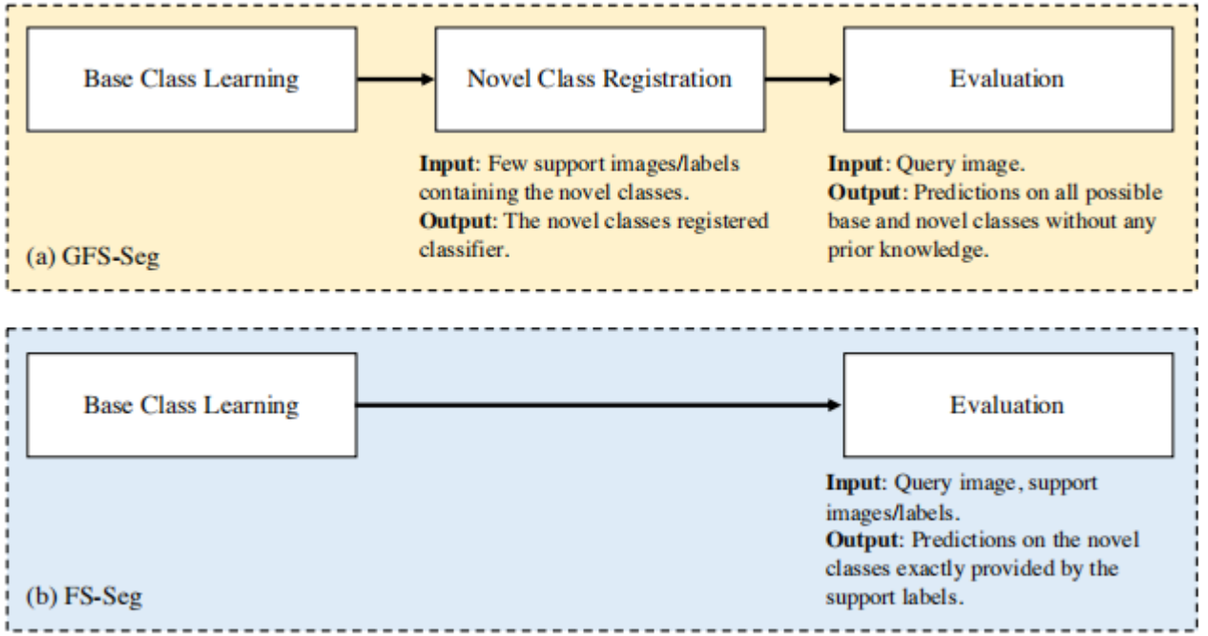
[1]CUHK [2]MPI Informatics [3]SmartMore [4]Cornell University [5]HKU [6]MIT

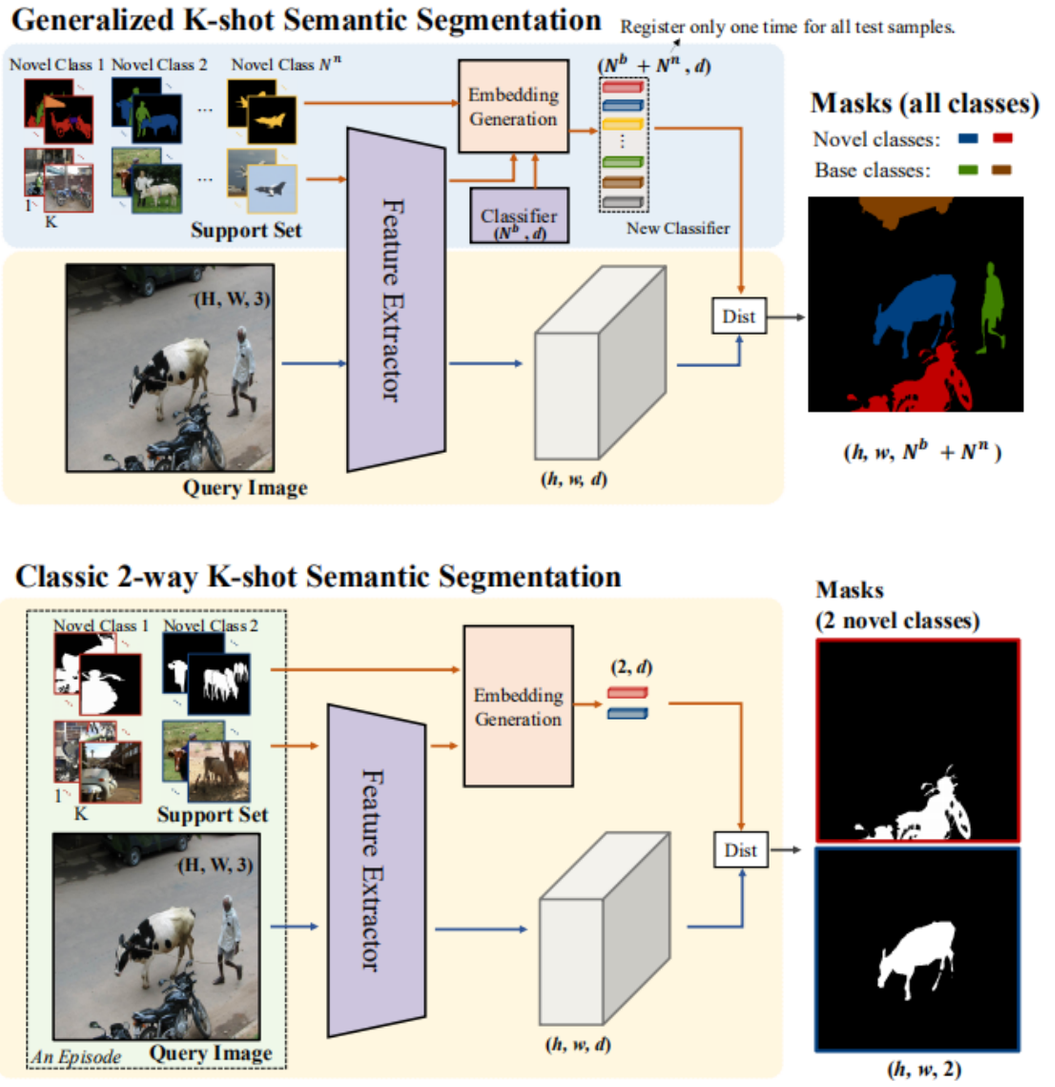
- **Limitations of Few-Shot-Seg**
  - FS-Seg requires support samples to contain classes that exist in query samples.
  - FS-Seg only evaluates the novel classes, while test samples in normal semantic segmentation may also contain the base classes.

- **Introduce a new benchmark, called Generalized Few-Shot Semantic Segmentation**

02 Generalized Few-shot Semantic Segmentation

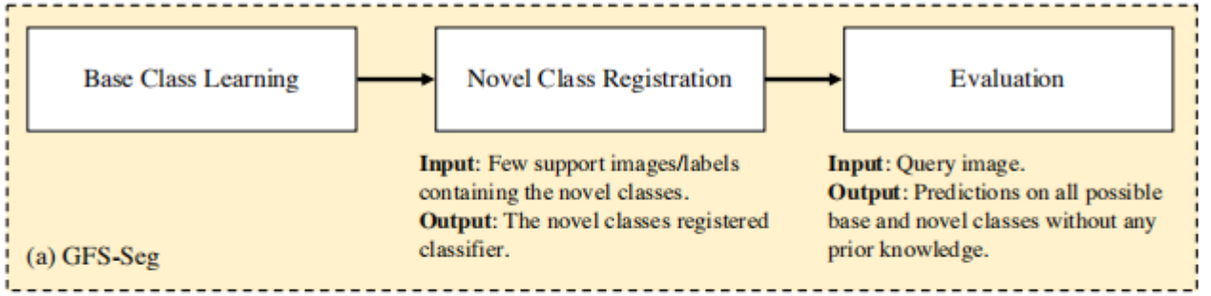
- **Difference between 2 settings**

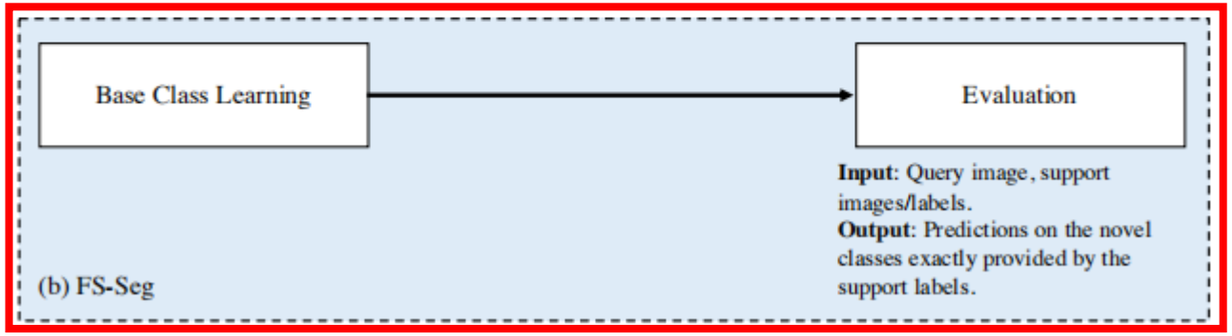


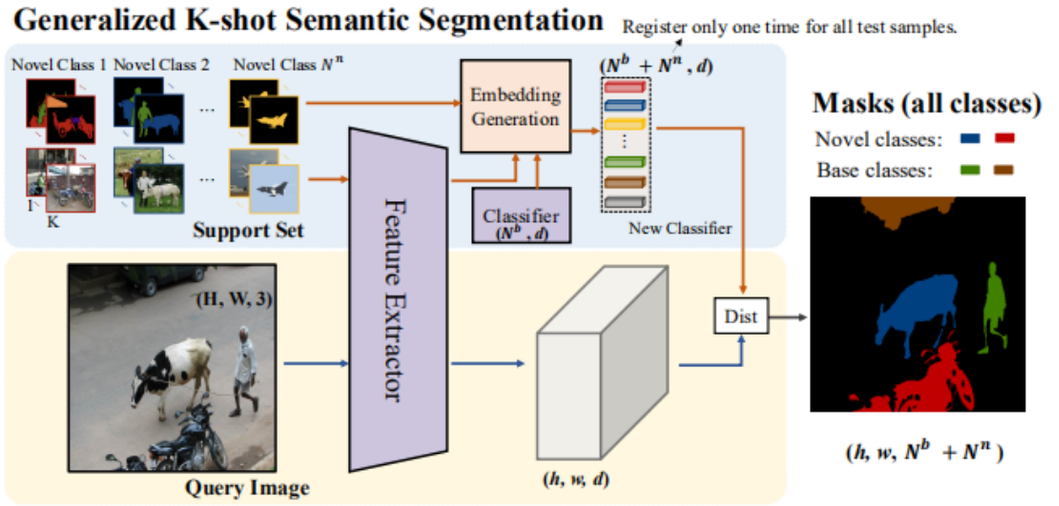


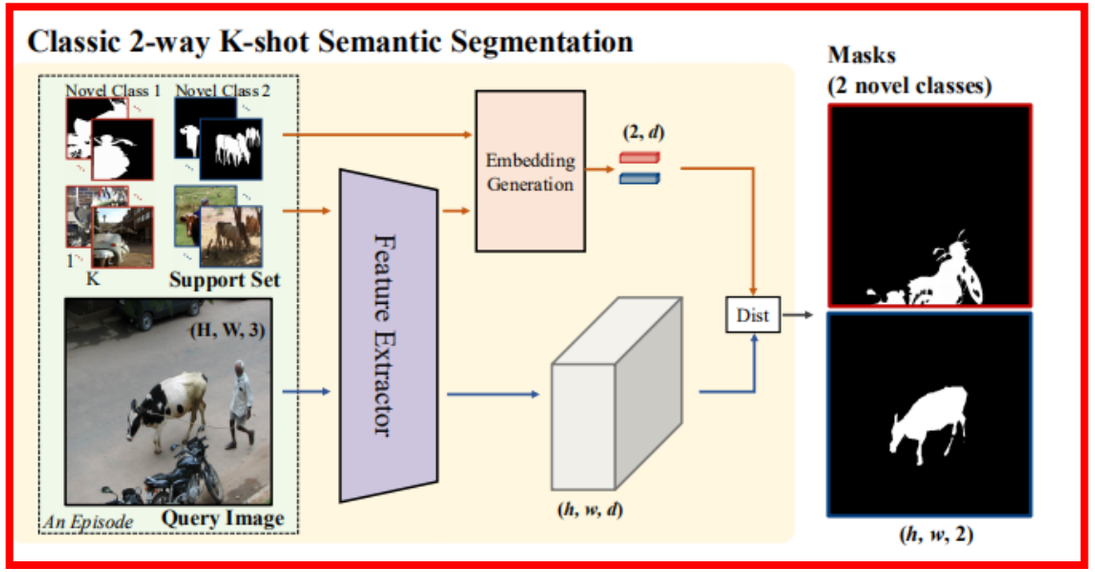
02 Generalized Few-shot Semantic Segmentation

- **Difference between 2 settings**



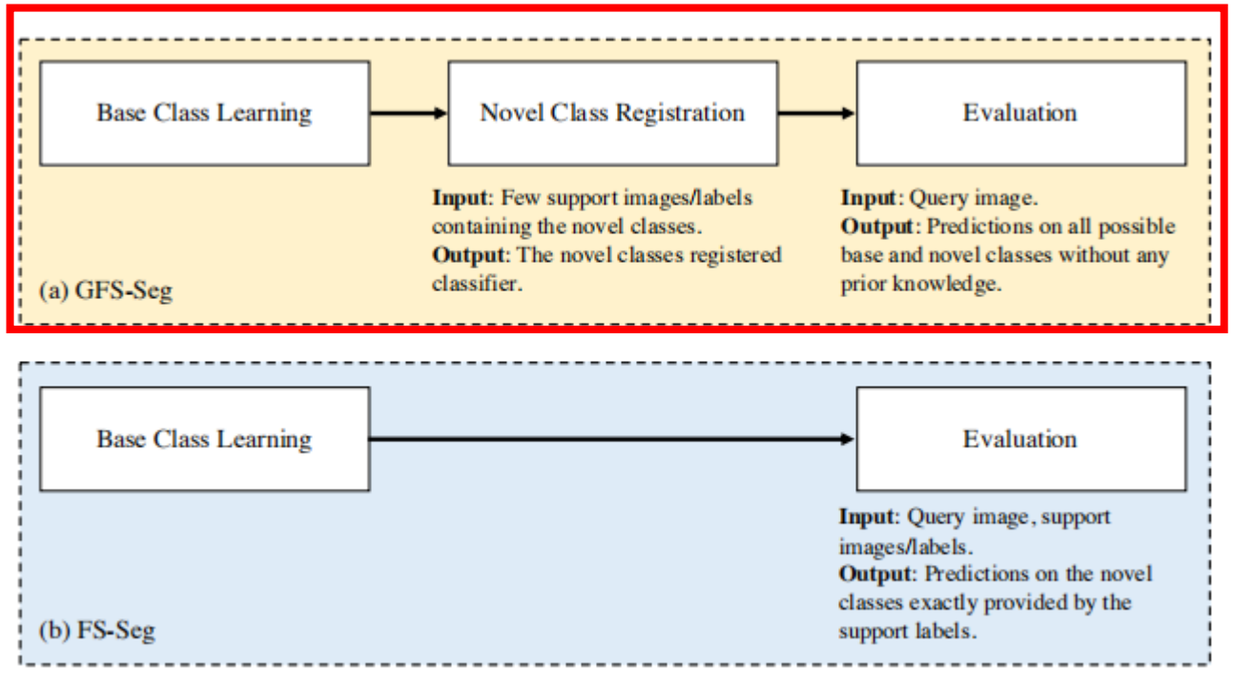


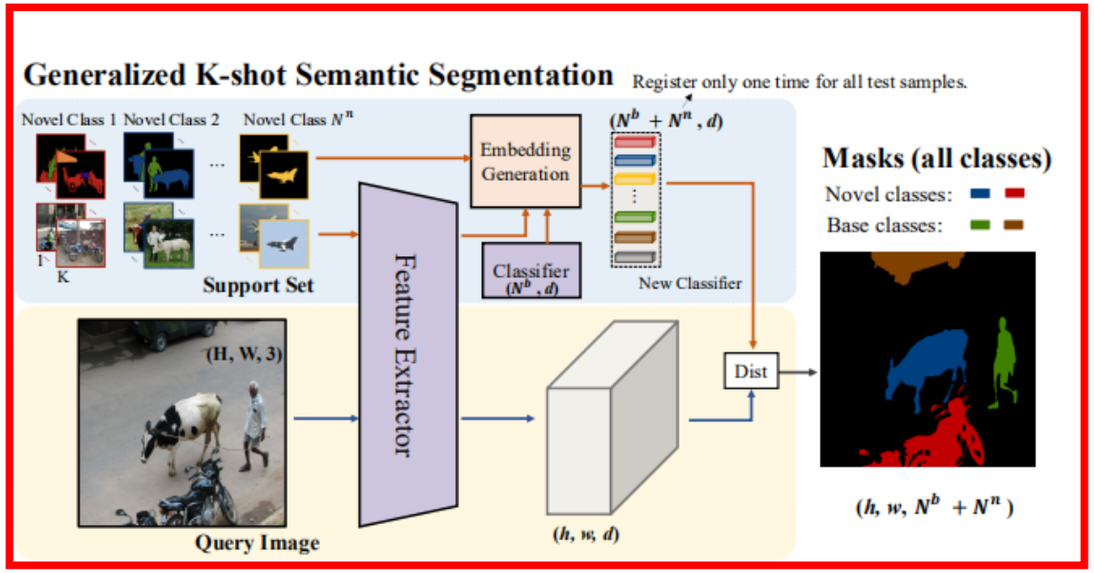


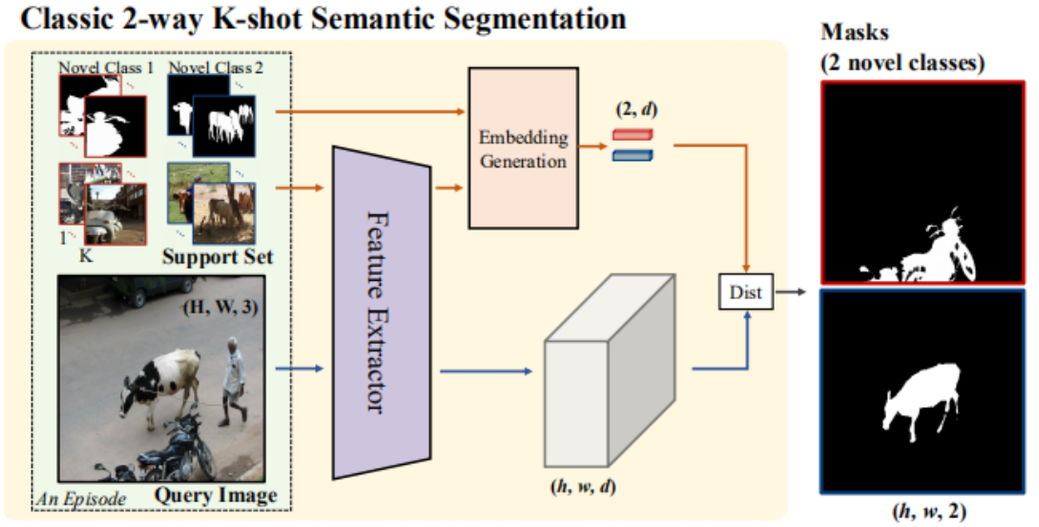


02 Generalized Few-shot Semantic Segmentation

- **Difference between 2 settings**







02 Generalized Few-shot Semantic Segmentation

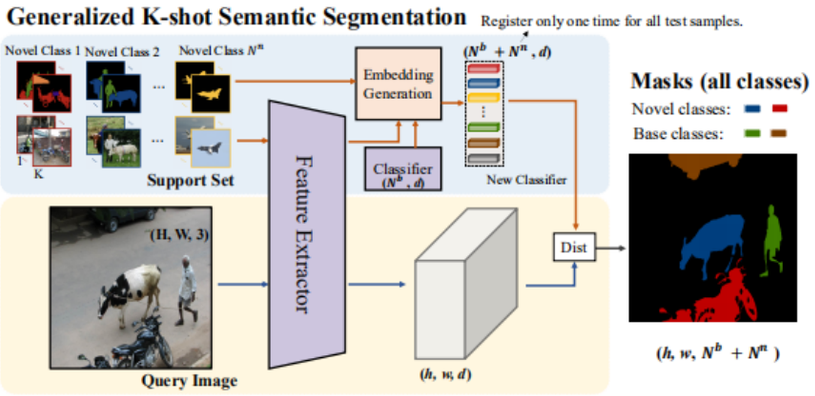
- **The baseline for GFS-Seg.**

- $\boldsymbol{P^n}$ $$\boldsymbol{p}^i = \frac{1}{K} * \sum_{j=1}^{K} \frac{\sum_{h,w}[\boldsymbol{m}_j^i \circ \mathcal{F}(\boldsymbol{s}_j^i)]_{h,w}}{\sum_{h,w}[\boldsymbol{m}_j^i]_{h,w}}, \qquad \bar{\boldsymbol{P}}^n \in \mathbb{R}^{N^n,d}$$

- $\boldsymbol{P^b}$ The common semantic segmentation frameworks（*e.g.* FCN, DeepLab, PSPNet） can be decomposed into two parts: feature extractor and classifier. The classifier of size $N^b \times$ d can be seen as N^b base prototypes

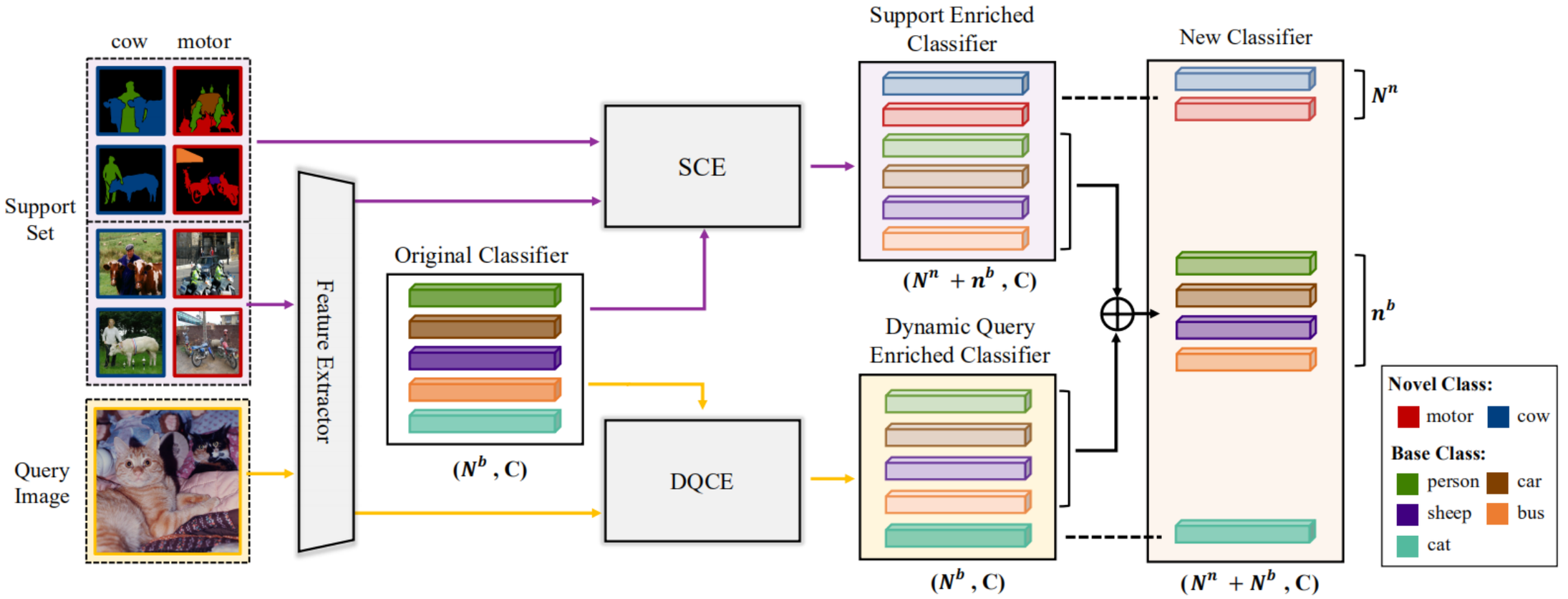
$$\boldsymbol{P}^b \in \mathbb{R}^{N^b,d}$$

- $\boldsymbol{P^{all}}$ $$\boldsymbol{P}^{all} \in \mathbb{R}^{N^b+N^n,d},$$



02 Generalized Few-shot Semantic Segmentation

- **Framework**
  - The weights of N^n novel classes are directly set by the averaged novel features.
  - Weights of n^b base classes appear in support samples are enriched by **SCE** with the original weights.
  - **DQCE** dynamically enriches the weights of N^b base classes from query samples



02 Generalized Few-shot Semantic Segmentation

- **Support Contextual Enrichment (SCE)**

$$\boldsymbol{p}^{b,i} = \boldsymbol{\gamma}^{i}_{sup} * \boldsymbol{p}^{b,i}_{cls} + (1 - \boldsymbol{\gamma}^{i}_{sup}) * \boldsymbol{p}^{b,i}_{sup}, \quad i \in \{1, ..., n^b\},$$

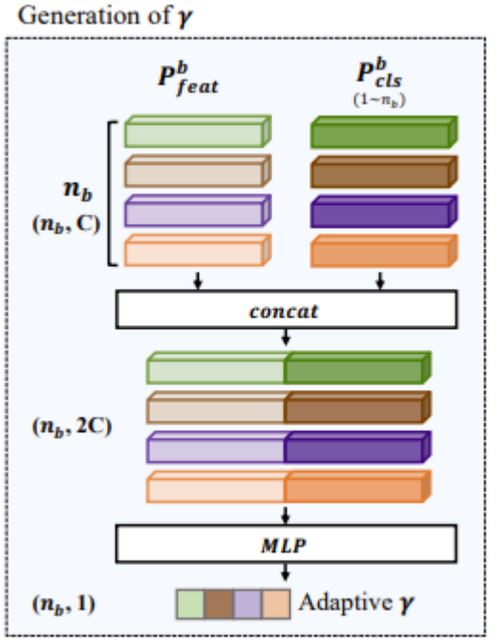
- **Dynamic Query Contextual Enrichment (DQCE)**

$$\boldsymbol{p}^{b}_{qry} = Softmax(\boldsymbol{y}^{t}_{qry}) \times \mathcal{F}(\boldsymbol{q}), \quad \boldsymbol{y}_{qry} \in \mathbb{R}^{h_q w_q \times N^b}$$

$$\boldsymbol{p}^{b,i}_{dyn} = \boldsymbol{\gamma}^{i}_{qry} * \boldsymbol{p}^{b,i}_{cls} + (1 - \boldsymbol{\gamma}^{i}_{qry}) * \boldsymbol{p}^{b,i}_{qry}, \quad i \in \{1, ..., N^b\}.$$

- **Final prototypes**

$$\boldsymbol{p}^{b,i}_{capl} = \boldsymbol{p}^{b,i} + \boldsymbol{p}^{b,i}_{dyn}, \quad i \in \{1, ..., N^b\}.$$



Why update the prototypes of the base classes?

types only capture the contextual relation between *Dog* and *People* during training. If *Sofa* is a novel class and some instances of *Sofa* in support samples appear with *Dog* (*e.g.*, a dog is lying on the sofa), merely mask-pooling in each support sample of *Sofa* to form the novel prototype may result in the base prototype of *Dog* losing the contextual co-occurrence information with *Sofa* and hence yield inferior results. Thus, for GFS-Seg, reasonable utilization of con-

02 Generalized Few-shot Semantic Segmentation

- **Results**

Methods	1-shot			5-shot		
	Base	Novel	Total	Base	Novel	Total
CANet [58]	8.73	2.42	7.23	9.05	1.52	7.26
PFENet [39]	8.32	2.67	6.97	8.83	1.89	7.18
SCL [58]	8.88	2.44	7.35	9.11	1.83	7.38
PANet [43]	31.88	11.25	26.97	32.95	15.25	28.74
PANet + CAPL	63.06	14.96	51.60	63.81	19.66	53.30
DeepLab-V3 + CAPL	65.71	15.05	53.77	67.01	23.26	56.59
PSPNet + CAPL	65.48	18.85	54.38	66.14	22.41	55.72

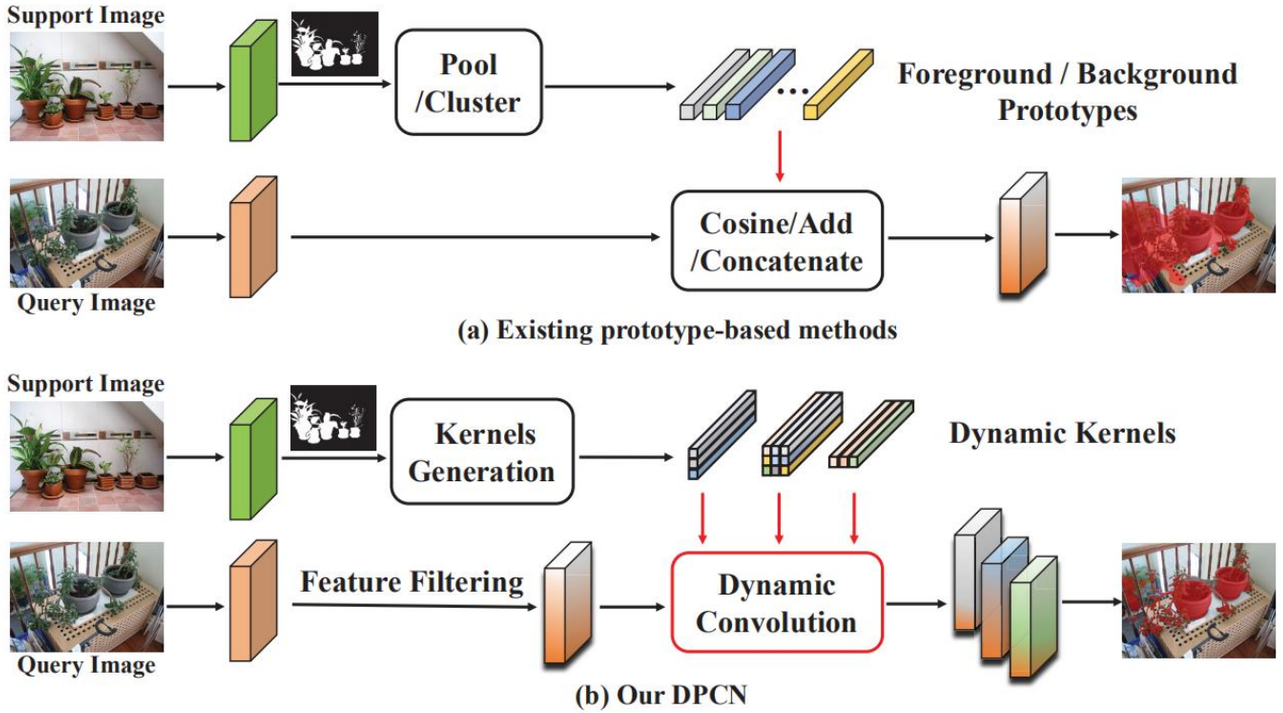
Methods	Venue	Backbone	Pascal-5^i		COCO-20^i	
			1-Shot	5-Shot	1-Shot	5-Shot
PANet [43]	ICCV-19	Res-50	48.1	55.7	20.9	29.7
PFENet [39]	TPAMI-20	Res-50	60.8	61.9	32.1	37.5
ASGNet [17]	CVPR-21	Res-50	59.3	63.9	34.5	42.5
SCL [55]	CVPR-21	Res-50	61.8	62.9	-	-
SAGNN [46]	CVPR-21	Res-50	62.1	62.8	-	-
RePri [3]	CVPR-21	Res-50	59.1	66.8	34.0	42.1
CWT [22]	ICCV-21	Res-50	56.4	63.7	32.9	41.3
MMNet [45]	ICCV-21	Res-50	61.8	63.4	37.5	38.2
CMN [47]	ICCV-21	Res-50	62.8	63.7	39.3	43.1
Mining [49]	ICCV-21	Res-50	62.1	66.1	33.9	40.6
HSNet [23]	ICCV-21	Res-50	**64.0**	**69.5**	39.2	46.9
CAPL (PANet)		Res-50	60.6	66.1	38.0	47.3
CAPL (PFENet)		Res-50	62.2	67.1	**39.8**	**48.3**
PFENet [39]	TPAMI-20	Res-101	60.1	61.4	32.4	37.4
SAGNN [46]	CVPR-21	Res-101	-	-	37.2	42.7
ASGNet [17]	CVPR-21	Res-101	59.3	64.4	-	-
CWT [22]	ICCV-21	Res-101	58.0	64.7	32.4	42.0
Mining [49]	ICCV-21	Res-101	62.6	68.8	36.4	44.4
HSNet [23]	ICCV-21	Res-101	**66.2**	**70.4**	41.2	49.5
CAPL (PFENet)		Res-101	63.6	68.9	**42.8**	**50.4**

03 Dynamic Prototype Convolution Network for Few-Shot Semantic Segmentation

Jie Liu[1*], Yanqi Bao[2*], Guo-Sen Xie[3,4†], Huan Xiong[4], Jan-Jakob Sonke[5], Efstratios Gavves[1]
[1]University of Amsterdam, Netherlands [2]Northeastern University, China [5]The Netherlands Cancer Institute, Netherlands
[3]Nanjing University of Science and Technology, China [4]Mohamed bin Zayed University of Artificial Intelligence, UAE

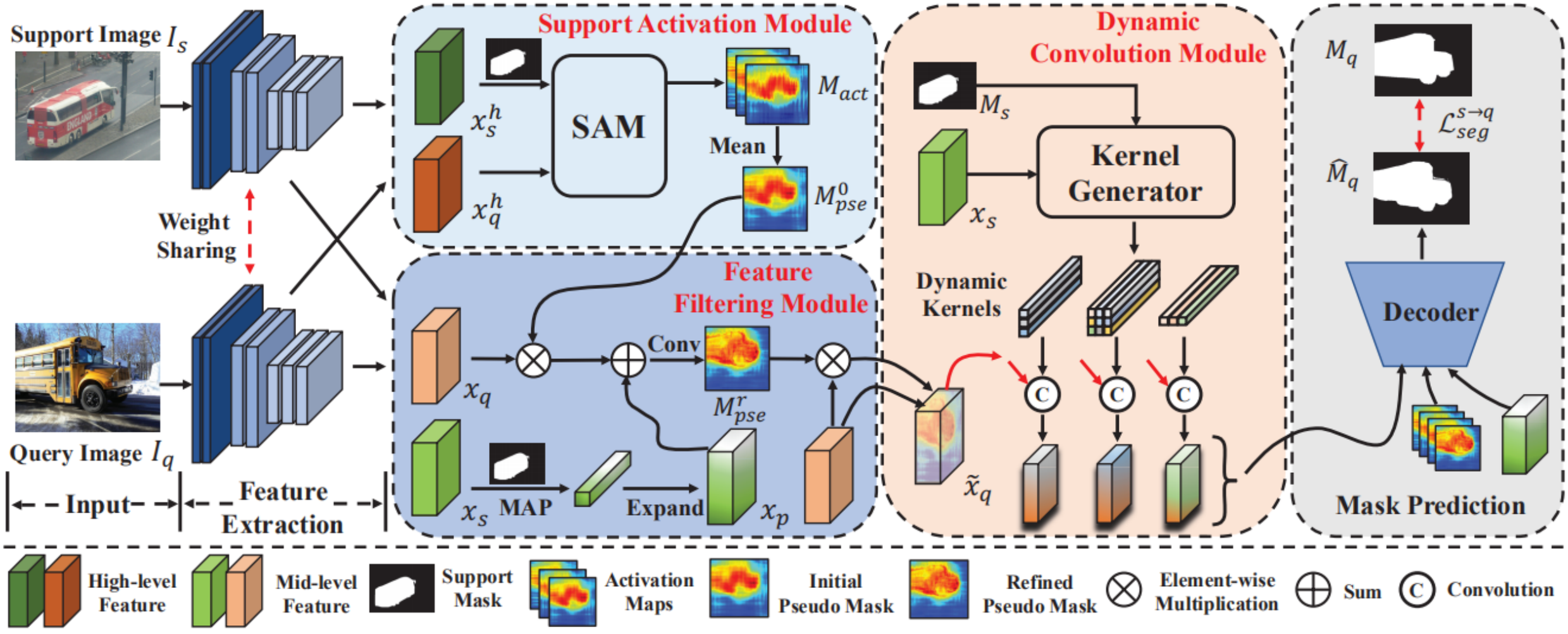
Most existing FSS methods usually cannot well capture the intrinsic object details in the query images that are widely encountered in FSS

DPCN can well capture the intrinsic subtle details. This benefits from dynamic convolution on query features with dynamic kernels generated from the support foreground features.



03 Dynamic Prototype Convolution Network for Few-Shot Semantic Segmentation

- **Framework**

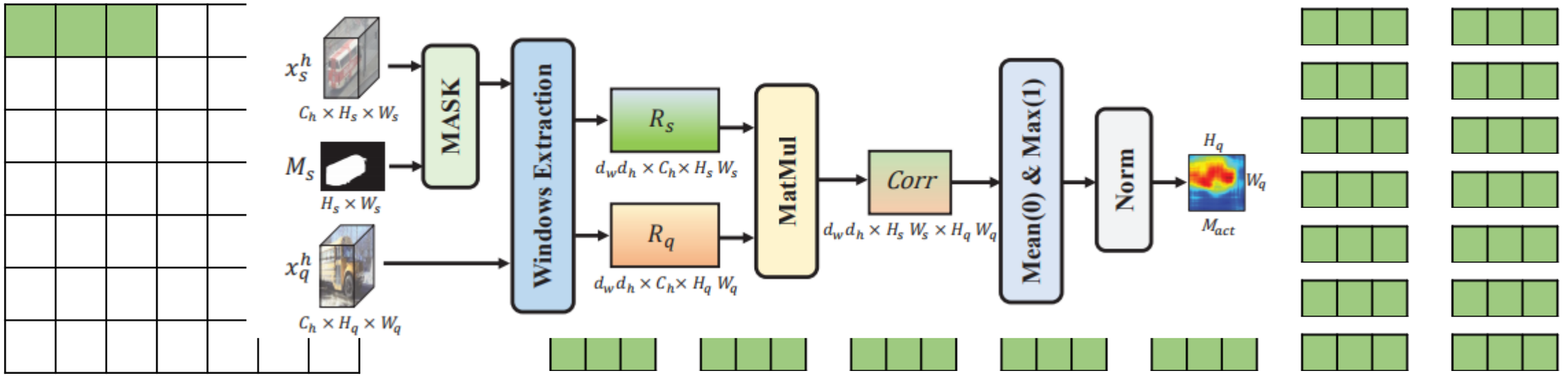


03 Dynamic Prototype Convolution Network for Few-Shot Semantic Segmentation

- **Support Activation Module (SAM)**

$$R_s = \mathcal{W}(x_s^h \otimes M_s) \in \mathbb{R}^{d_h d_w \times C_h \times H_s W_s},$$

$$R_q = \mathcal{W}(x_q^h) \in \mathbb{R}^{d_h d_w \times C_h \times H_q W_q},$$



Feature 7×7

Window size 3×1

$R \in 3 \times 1 \times 7 \times 7$

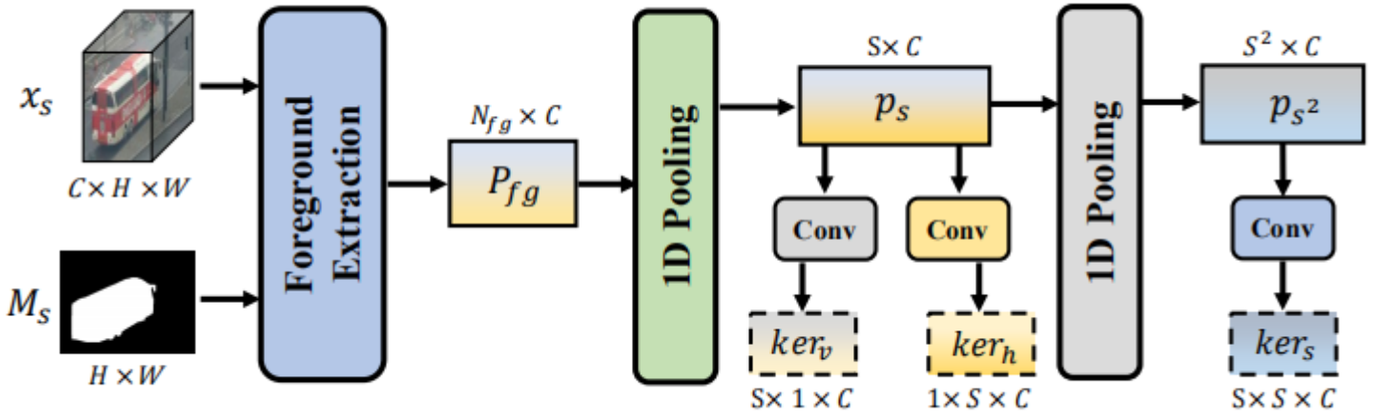
Regional matching map: $Corr \in \mathbb{R}^{d_h d_w \times H_s W_s \times H_q W_q}$.

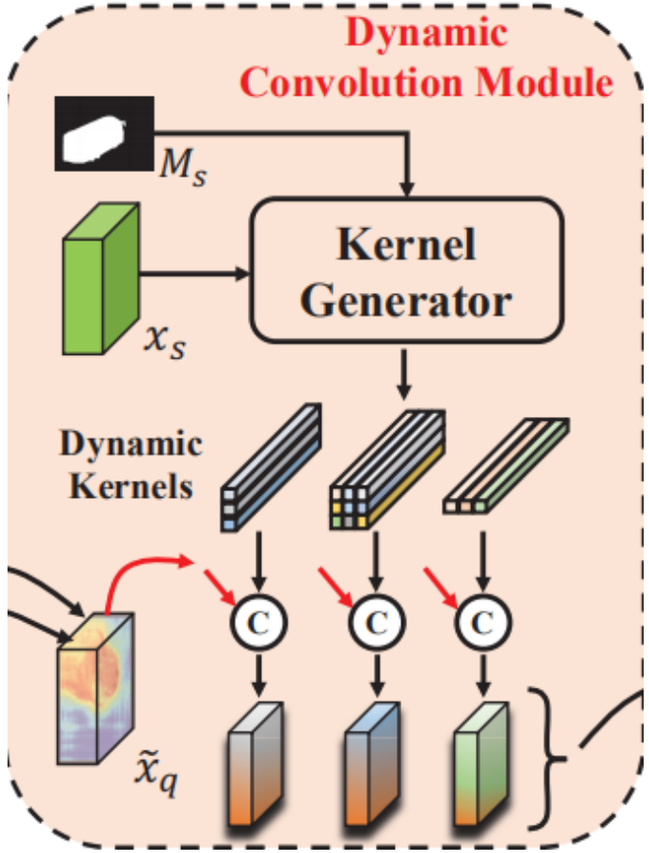
03 Dynamic Prototype Convolution Network for Few-Shot Semantic Segmentation

- **Dynamic Convolution Module**

$$P_{fg} = \mathcal{F}_e(x_s \otimes M_s) \in \mathbb{R}^{N_{fg} \times C},$$

$$p_s = pool_s(P_{fg}), p_{s^2} = pool_{s^2}(p_s).$$





03 Dynamic Prototype Convolution Network for Few-Shot Semantic Segmentation

- **Results**

Methods	Backbone	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
OSLSM (BMVC'17) [20]	VGG16	33.6	55.3	40.9	33.5	40.8	61.3	35.9	58.1	42.7	39.1	43.9	61.5
co-FCN (ICLRW'18) [19]	VGG16	36.7	50.6	44.9	32.4	41.1	60.1	37.5	50.0	44.1	33.9	41.4	60.2
AMP-2(ICCV'19) [21]	VGG16	41.9	50.2	46.7	34.7	43.4	61.9	40.3	55.3	49.9	40.1	46.4	62.1
PFENet (TPAMI'20) [25]	VGG16	56.9	68.2	54.4	52.4	58.0	72.0	59.0	69.1	54.8	52.9	59.0	72.3
HSNet (ICCV'21) [17]	VGG16	59.6	65.7	59.6	54.0	59.7	73.4	64.9	69.0	64.1	58.6	64.1	76.6
PFENet (TPAMI'20) [25]	ResNet50	61.7	69.5	55.4	56.3	60.8	73.3	63.1	70.7	55.8	57.9	61.9	73.9
RePRI (CVPR'21) [1]	ResNet50	59.8	68.3	62.1	48.5	59.7	-	64.6	71.4	**71.1**	59.3	66.6	-
SAGNN (CVPR'21) [30]	ResNet50	64.7	69.6	57.0	57.3	62.1	73.2	64.9	70.0	57.0	59.3	62.8	73.3
SCL (CVPR'21) [35]	ResNet50	63.0	70.0	56.5	57.7	61.8	71.9	64.5	70.9	57.3	58.7	62.9	72.8
MLC (ICCV'21) [32]	ResNet50	59.2	71.2	65.6	52.5	62.1	-	63.5	71.6	71.2	58.1	66.1	-
MMNet (ICCV'21) [29]	ResNet50	62.7	70.2	57.3	57.0	61.8	-	62.2	71.5	57.5	62.4	63.4	-
HSNet (ICCV'21) [17]	ResNet50	64.3	70.7	60.3	60.5	64.0	76.7	**70.3**	**73.2**	67.4	**67.1**	69.5	80.6
Baseline	VGG16	58.4	68.0	58.0	50.9	58.8	71.2	60.7	68.8	60.2	52.2	60.4	74.3
DPCN	VGG16	58.9	69.1	63.2	55.7	61.7	73.7	63.4	70.7	68.1	59.0	65.3	77.2
Baseline	ResNet50	61.1	69.8	58.4	56.3	61.4	71.5	63.7	70.9	58.7	57.4	62.7	73.7
DPCN	ResNet50	**65.7**	**71.6**	**69.1**	**60.6**	**66.7**	**78.0**	70.0	**73.2**	70.9	65.5	**69.9**	**80.7**

Table 1. Comparison with state-of-the-arts on PASCAL-5^i dataset under both 1-shot and 5-shot settings. mIoU of each fold, and averaged mIoU & FB-IoU of all folds are reported. Baseline results are achieved by removing three modules (i.e., SAM, FFM, and DCM) in DPCN.

Methods	Backbone	1-shot						5-shot					
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU	Fold-0	Fold-1	Fold-2	Fold-3	Mean	FB-IoU
FWB(ICCV'19) [18]	VGG16	18.4	16.7	19.6	25.4	20.0	-	20.9	19.2	21.9	28.4	22.6	-
PFENet(TPAMI'20) [25]	VGG16	33.4	36.0	34.1	32.8	34.1	60.0	35.9	40.7	38.1	36.1	37.7	61.6
SAGNN(CVPR'21) [30]	VGG16	35.0	40.5	37.6	36.0	37.3	61.2	37.2	45.2	40.4	40.0	40.7	63.1
RePRI(CVPR'21) [1]	ResNet50	31.2	38.1	33.3	33.0	34.0	-	38.5	46.2	40.0	43.6	42.1	-
MLC(ICCV'21) [32]	ResNet50	**46.8**	35.3	26.2	27.1	33.9	-	**54.1**	41.2	34.1	33.1	40.6	-
MMNet(ICCV'21) [29]	ResNet50	34.9	41.0	37.2	37.0	37.5	-	37.0	40.3	39.3	36.0	38.2	-
HSNet(ICCV'21) [17]	ResNet50	36.3	43.1	38.7	38.7	39.2	**68.2**	43.3	51.3	48.2	45.0	46.9	**70.7**
SAGNN(CVPR'21) [30]	ResNet101	36.1	41.0	38.2	33.5	37.2	60.9	40.9	48.3	42.6	38.9	42.7	63.4
SCL(CVPR'21) [35]	ResNet101	36.4	38.6	37.5	35.4	37.0	-	38.9	40.5	41.5	38.7	39.9	-
Baseline	VGG16	32.1	36.1	35.2	32.3	33.9	60.1	35.0	40.1	37.1	36.5	37.2	61.8
DPCN	VGG16	38.5	43.7	38.2	37.7	39.5	62.5	42.7	51.6	45.7	44.6	46.2	66.1
Baseline	ResNet50	32.3	38.3	34.9	32.5	34.5	57.7	35.0	41.0	37.3	35.5	37.2	59.2
DPCN	ResNet50	42.0	**47.0**	**43.2**	**39.7**	**43.0**	63.2	46.0	**54.9**	**50.8**	**47.4**	**49.8**	67.4

Table 2. Comparison with state-of-the-arts on COCO-20^i dataset under both 1-shot and 5-shot settings. mIoU of each fold, and averaged mIoU & FB-IoU of all folds are reported. Baseline results are achieved by removing three modules (i.e., SAM, FFM, and DCM) in DPCN.

Kernel Size	1-shot mIoU					FB-IoU
	Fold-0	Fold-1	Fold-2	Fold-3	Mean	
3	65.2	70.4	68.5	59.4	65.9	77.5
5	65.7	**71.6**	69.1	**60.6**	**66.7**	**78.0**
7	65.5	70.7	**69.3**	59.0	66.1	77.5
9	**65.9**	70.8	68.8	59.7	66.3	77.7

SAM	FFM	DCM	1-shot mIoU					FB-IoU
			Fold-0	Fold-1	Fold-2	Fold-3	Mean	
✓	✓		63.6	69.7	65.0	59.6	64.5	75.2
	✓	✓	**67.1**	71.1	63.2	60.0	65.4	76.0
✓		✓	63.5	70.8	65.7	58.9	64.7	75.8
✓	✓	✓	65.7	**71.6**	**69.1**	**60.6**	**66.7**	**78.0**

Methods	1-shot mIoU					FB-IoU
	Fold-0	Fold-1	Fold-2	Fold-3	Mean	
CANet	53.5	65.9	51.3	51.9	55.4	66.2
CANet+DCM	64.7	66.8	51.8	51.9	58.8	69.3
PFENet	61.7	69.5	55.4	56.3	60.8	73.3
PFENet+DCM	62.2	69.6	59.2	58.0	62.3	73.5

Table 6. Generalization ability of the proposed DCM.