Learning What Not to Segment: A New Perspective on Few-Shot Segmentation

Chunbo Lang Gong Cheng* Binfei Tu Junwei Han
School of Automation, Northwestern Polytechnical University, Xi'an, China
{langchunbo, binfeitu}@mail.nwpu.edu.cn, {gcheng, jhan}@nwpu.edu.cn

CVPR 2022

Presenter: Hao Wang

Advisor: Chia-Wen Lin

Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

Introduction

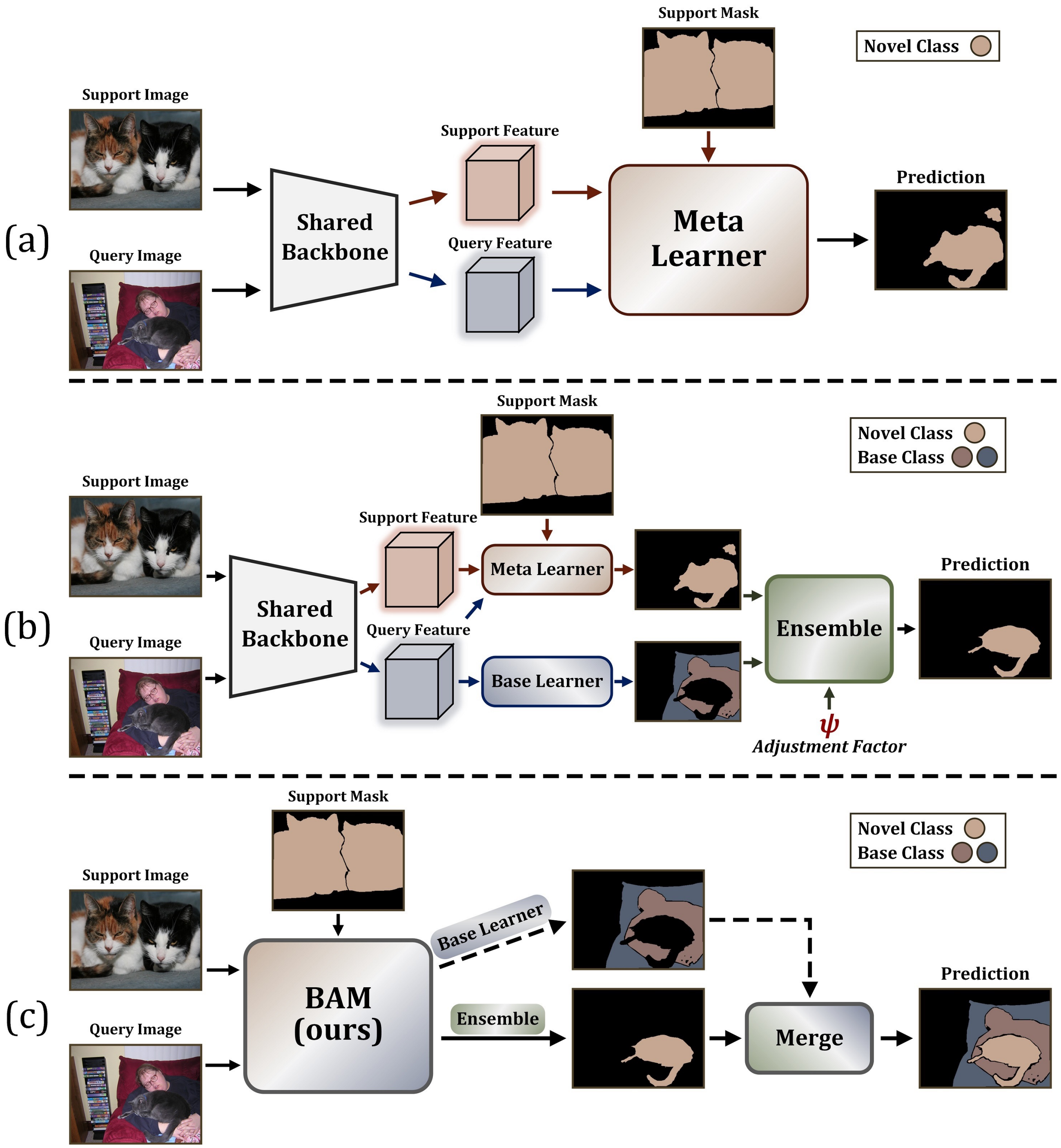
- Propose a simple but efficient scheme to address the bias problem by introducing **an additional branch to explicitly predict the regions of base classes** in the query images, which sheds light on future works.

- Propose to estimate the scene differences between the query-support image pairs through the **Gram matrix** for mitigating the adverse effects caused by the sensitivity of meta learner.

- Extend the proposed approach to a more challenging setting, **generalized FSS**, which simultaneously identifies the targets of base and novel classes.

Introduction

the proposed scheme is named BAM as it consists of two unique learners, base and the meta



Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

Related Work

- Pyramid Scene Parsing Network (PSPNet)
  - is served as the base learner
- Prior Guided Feature Enrichment Network for Few-Shot Segmentation(PFENet)
  - is served as the meta learner, where the FEM is replaced by ASPP

Related Work – PSPNet

Pyramid Scene Parsing Network

Hengshuang Zhao[1] Jianping Shi[2] Xiaojuan Qi[1] Xiaogang Wang[1] Jiaya Jia[1]
[1]The Chinese University of Hong Kong [2]SenseTime Group Limited
{hszhao, xjqi, leojia}@cse.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk, shijianping@sensetime.com

CVPR 2017

Related Work – PSPNet

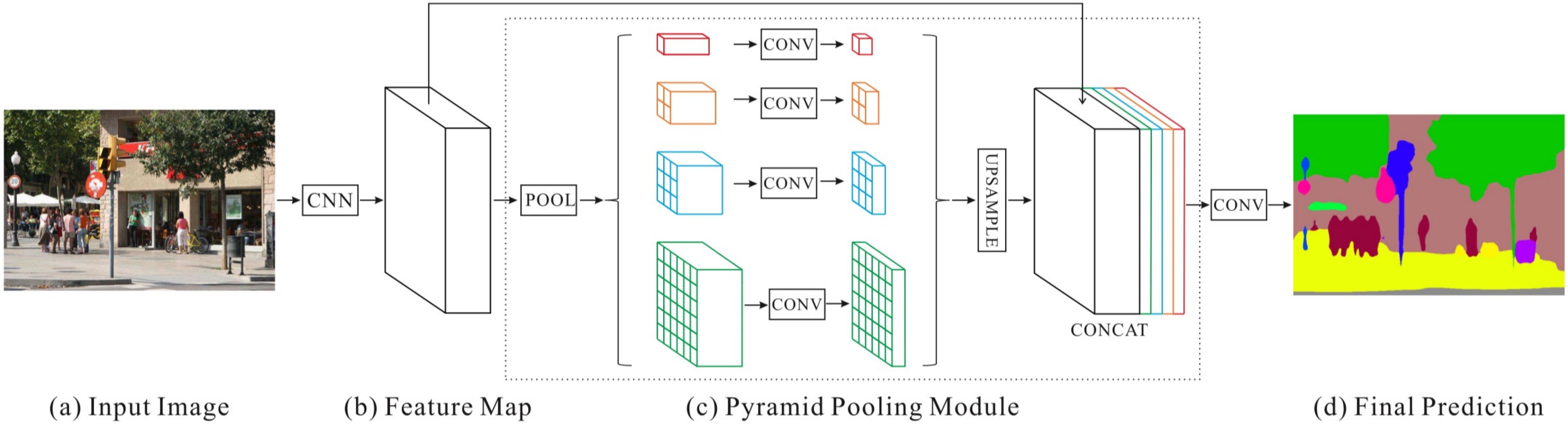
- Propose a **pyramid scene parsing network** to embed difficult scenery context features in an FCN based pixel prediction framework.

- Develop an **effective optimization strategy** for deep ResNet based on deeply supervised loss

Related Work – PSPNet

- **Mismatched Relationship**: There exist co-occurrent visual patterns. For example, an airplane is likely to be in runway or fly in sky while not over a road.
- **Confusion Categories**: dataset that are confusing in classification. Examples are wall, house, building and skyscraper.
- **Inconspicuous Classes**: Several small-size things, like streetlight and signboard, are hard to find.

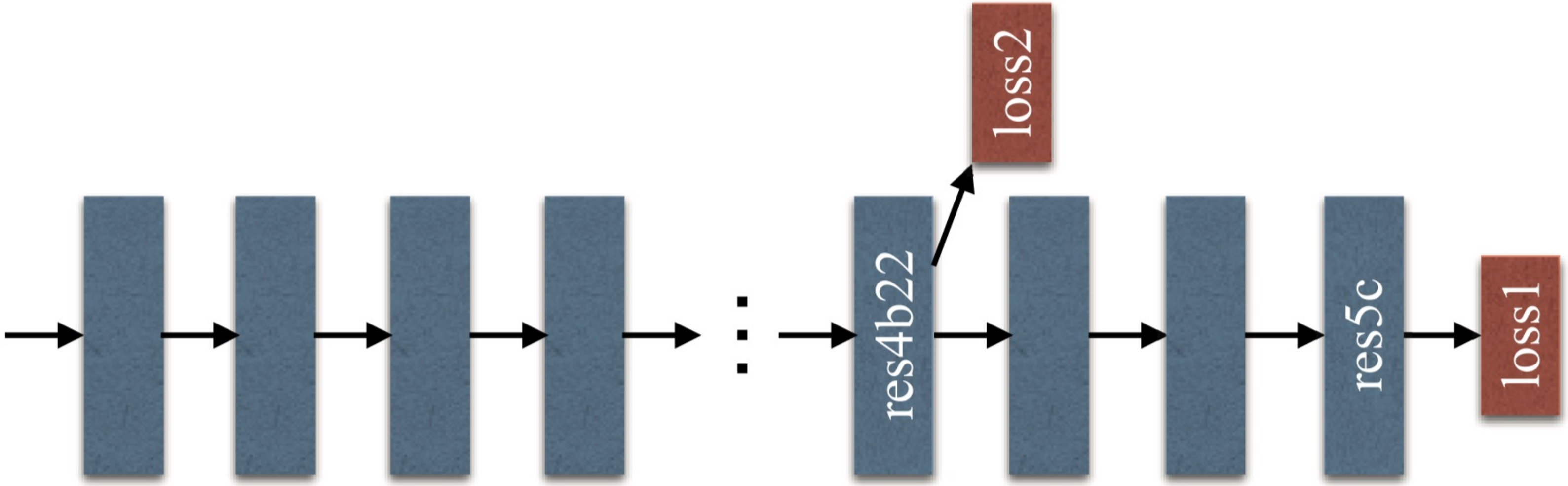
Related Work – PSPNet



- To summarize these observations, many errors related to **contextual relationship** and **global information** for different receptive fields.

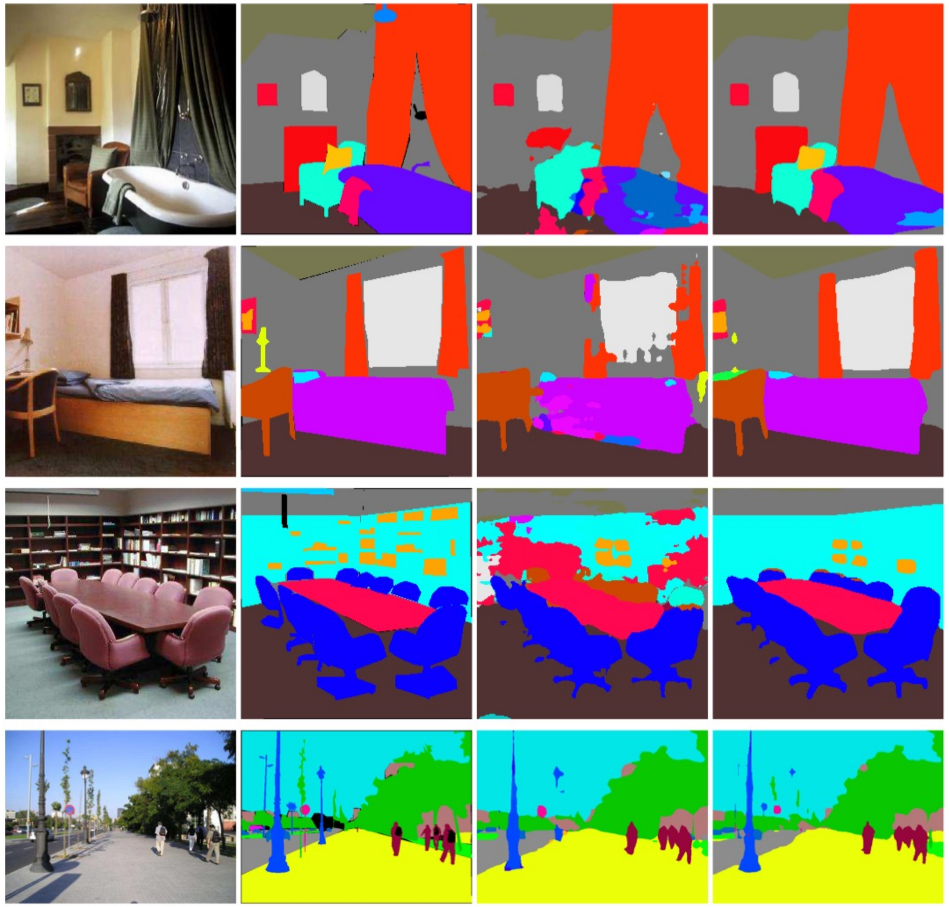
- A hierarchical global prior, containing information with different scales and varying among different sub-regions. Provides an **effective global contextual prior** for pixel-level scene parsing

Related Work – PSPNet



- Increasing depth of the network may introduce additional **optimization difficulty**
- Apart from the main branch using $softmax$ loss to train the final classifier, another classifier is applied after the fourth stage . The **auxiliary loss helps optimize the learning process**, while the master branch loss takes the most responsibility

Related Work – PSPNet



(a) Image (b) Ground Truth (c) Baseline (d) PSPNet

(a) Image (b) Ground Truth (c) Baseline (d) PSPNet

Figure 7. Visual improvements on PASCAL VOC 2012 data. PSPNet produces more accurate and detailed results.

- Our baseline network is FCN and dilated network

Related Work – PSPNet – Contribution

- Proposed an effective pyramid scene parsing network for complex scene understanding
- The global pyramid pooling feature provides **additional contextual information**
- Also provided a deeply **supervised optimization strategy** for ResNet-based FCN network

Related Work – PFENet

Prior Guided Feature Enrichment Network for Few-Shot Segmentation

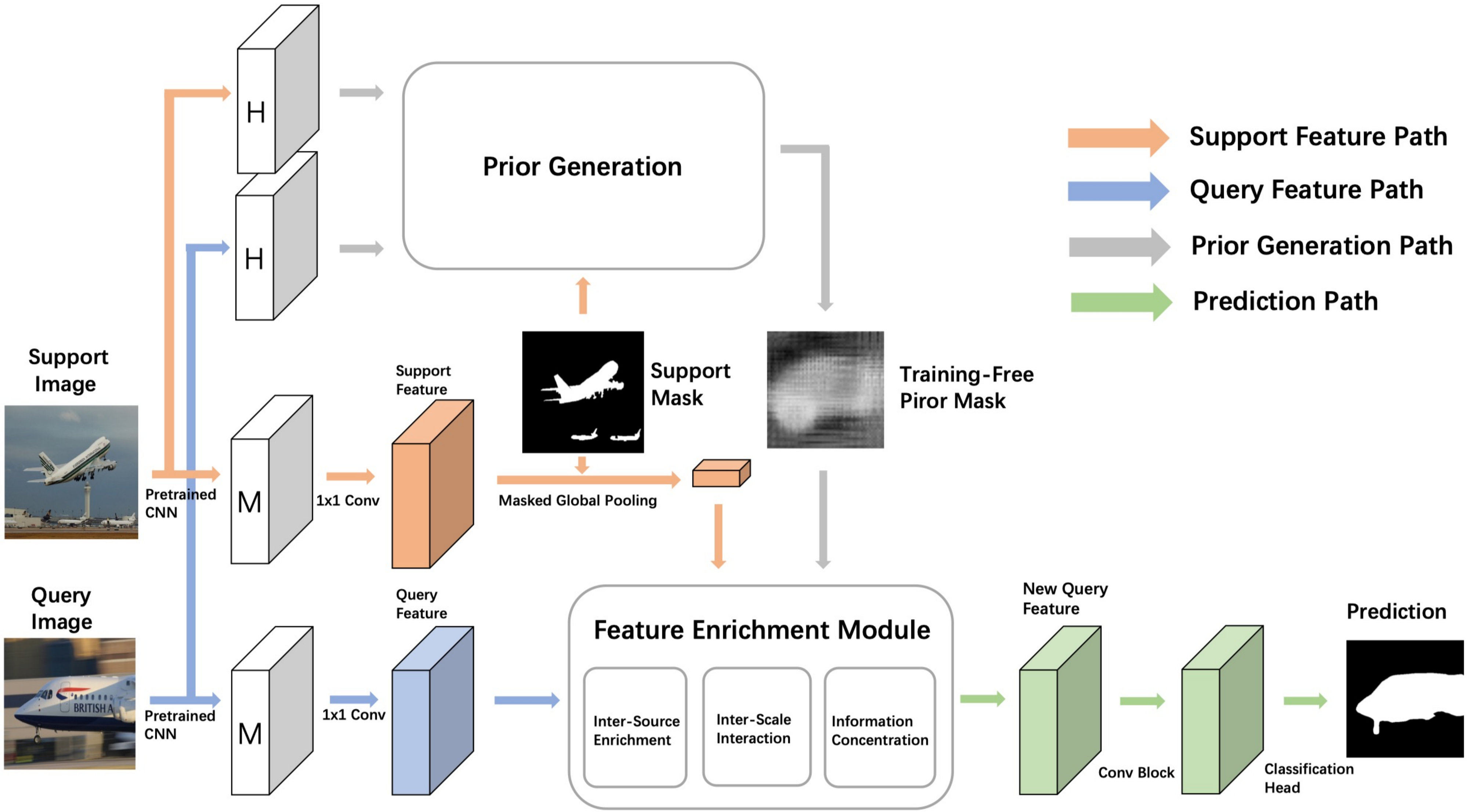
Zhuotao Tian, *Student Member, IEEE,* Hengshuang Zhao, *Member, IEEE,* Michelle Shu, *Student Member, IEEE,* Zhicheng Yang, *Member, IEEE,* Ruiyu Li, *Member, IEEE,* Jiaya Jia, *Fellow, IEEE*

TPAMI 2020

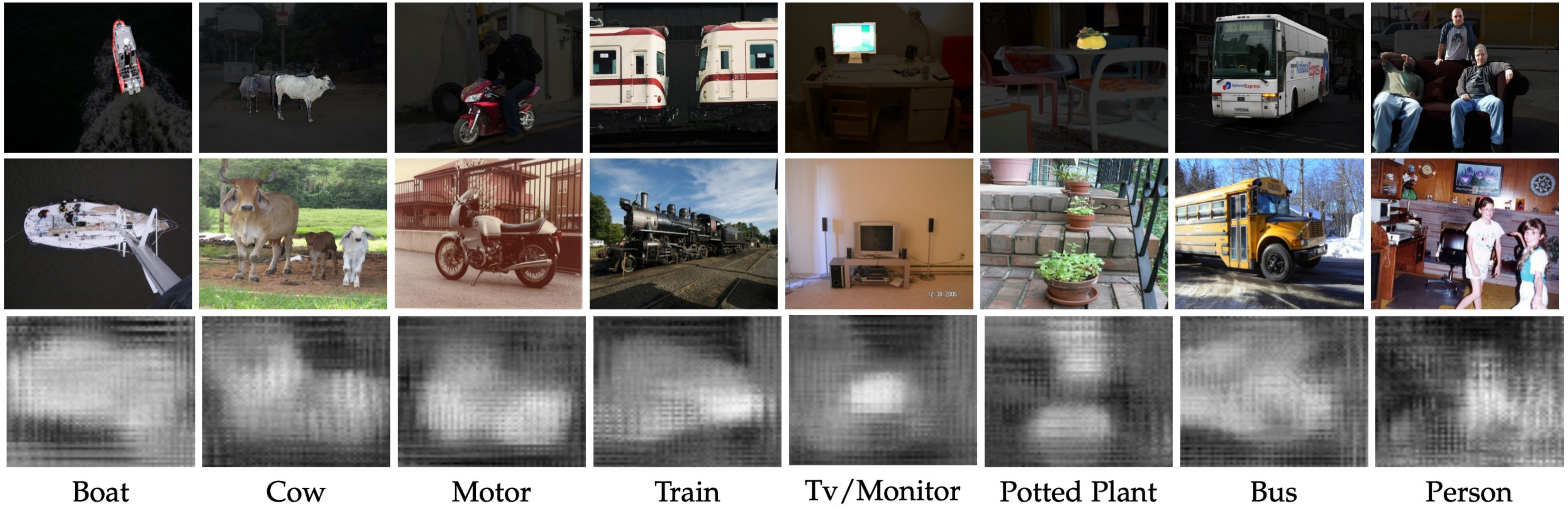
Related Work – PFENet

- Leverage high-level features and **propose training-free prior generation** to greatly improve prediction accuracy and retain high generalization
- By **incorporating the support feature** and **prior information**, our FEM helps adaptively **refine the query feature** with the conditioned inter-scale information interaction

Related Work – PFENet – Framework



Related Work – PFENet – Prior Generation



- High-level feature is **more class-specific** than the middle-level feature
- Exploit these features to **provide semantic cues** for final prediction

Related Work – PFENet – Prior Generation

$$X_Q = \mathcal{F}(I_Q), \quad X_S = \mathcal{F}(I_S) \odot M_S$$

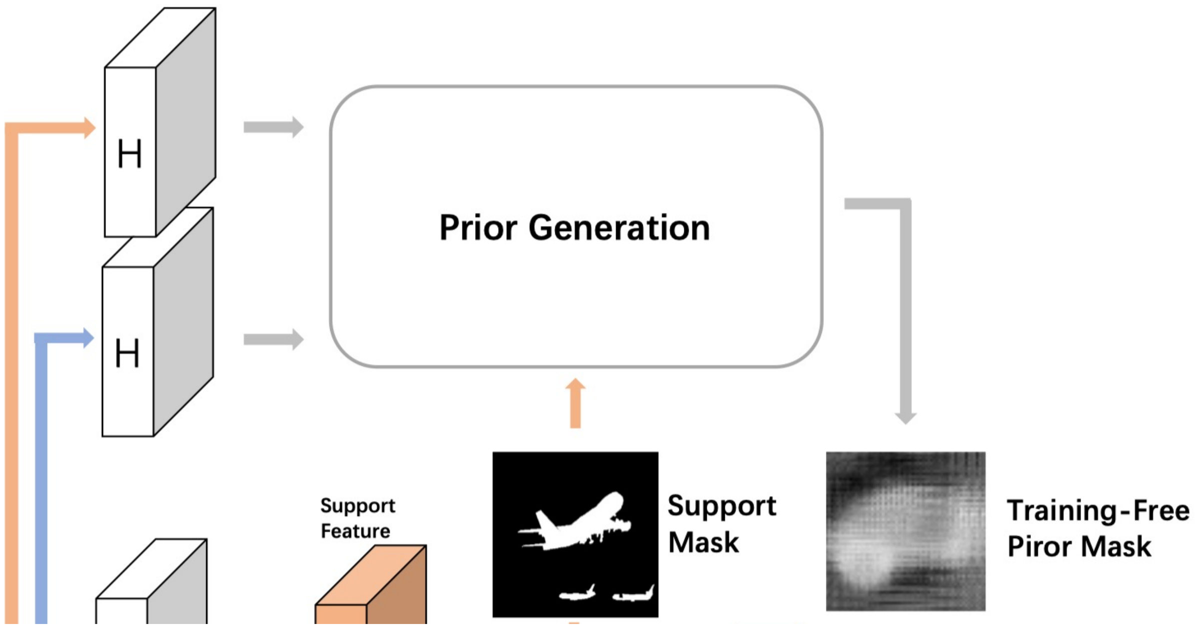
where $\odot$ is the Hadamard product

$$cos(x_q, x_s) = \frac{x_q^T x_s}{\|x_q\| \, \|x_s\|} \quad q, s \in \{1, 2, ..., hw\}$$

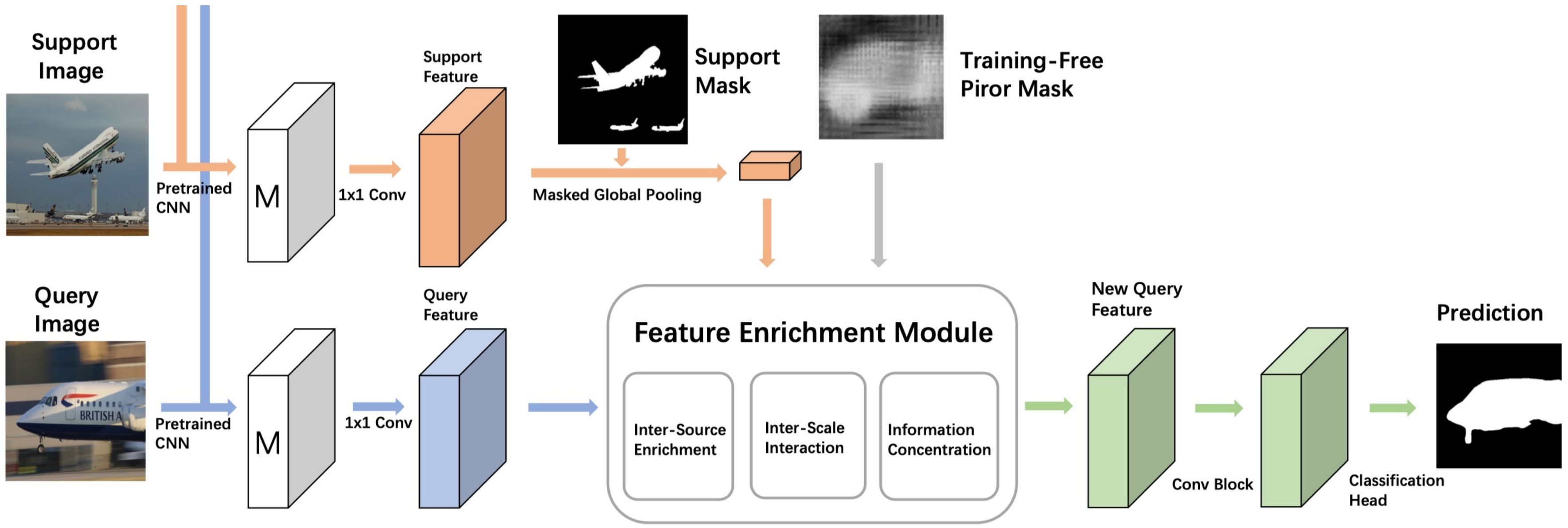
$$c_q = \max_{s \in \{1,2,...,hw\}} (cos(x_q, x_s)),$$
$$C_Q = [c_1, c_2, ..., c_{hw}] \in \mathbb{R}^{hw \times 1}.$$

Reshaping $C_Q \in \mathbb{R}^{hw*1}$ into $Y_Q \in \mathbb{R}^{h*w*1}$

$$Y_Q = \frac{Y_Q - \min(Y_Q)}{\max(Y_Q) - \min(Y_Q) + \epsilon}$$



Related Work – PFENet – Feature Enrichment Module



- **horizontally** interact the query feature with the support features and prior masks in each scale
- **vertically** leverage the hierarchical relations to enrich coarse feature maps with essential information extracted from the finer feature via a **top-down information** path
- features projected into different scales are then collected to **form the new query feature**

Related Work – PFENet – Feature Enrichment Module

- **Inter-Source Enrichment**

$$X_{Q,m}^{i} = \mathcal{F}_{1\times1}(X_Q^i \oplus X_S^i \oplus Y_Q^i)$$

n different spatial sizes

- **Inter-Scale Interaction**

$$X_{Q,new}^{i} = \mathcal{M}(X_{Q,m}^{Main,i}, X_{Q,m}^{Aux,i})$$

- **Information Concentration**

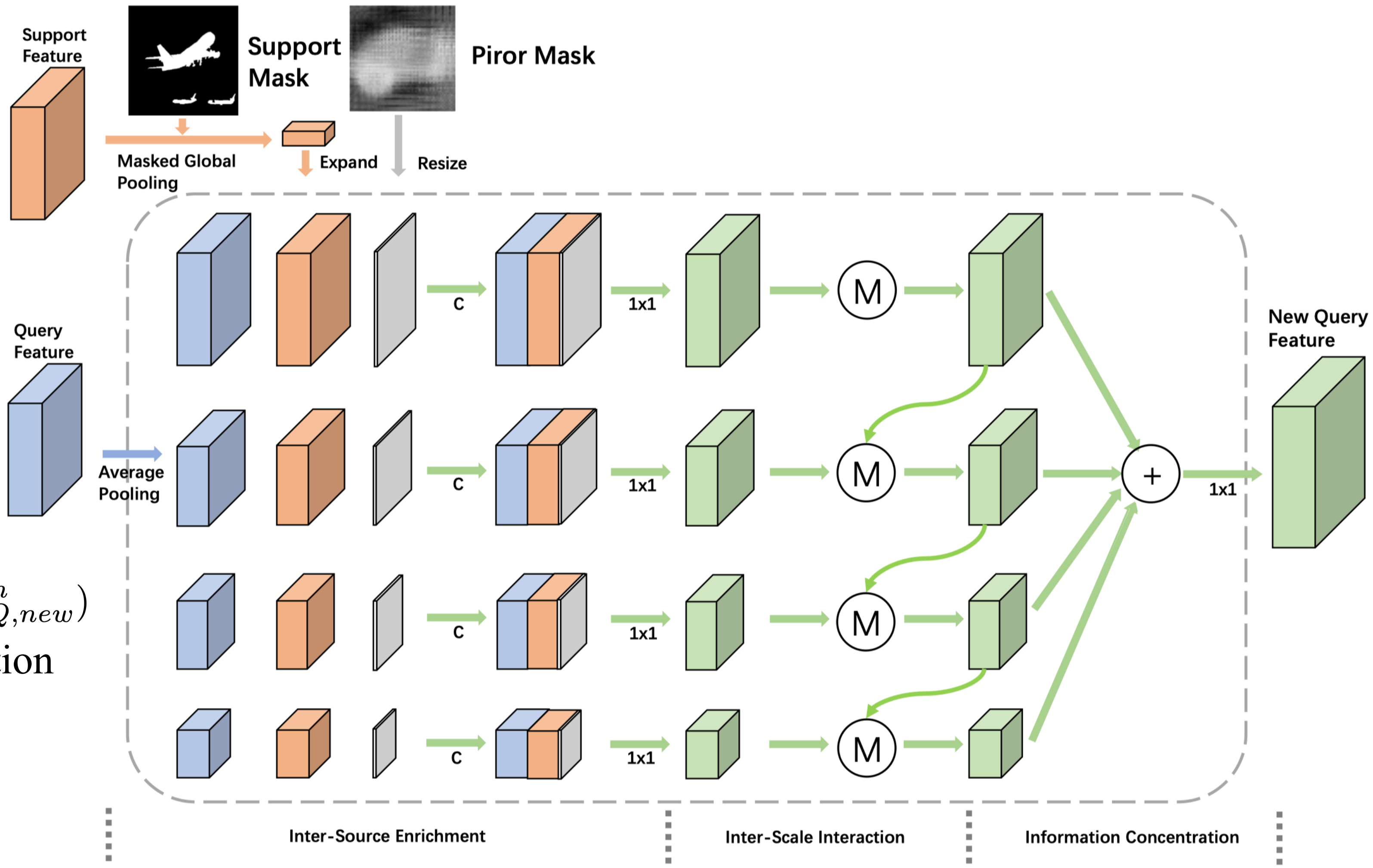
$$X_{Q,new} = \mathcal{F}_{1\times1}(X_{Q,new}^{1} \oplus X_{Q,new}^{2} \cdots \oplus X_{Q,new}^{n})$$

interpolation ,concatenation

- **Loss Function**

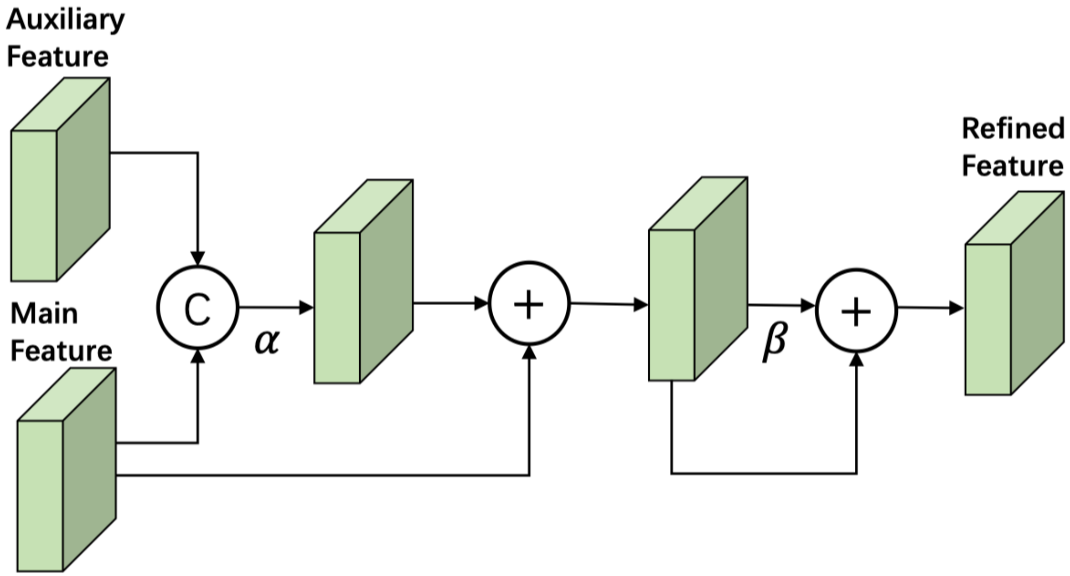
$$\mathcal{L} = \frac{\sigma}{n}\sum_{i=1}^{n}\mathcal{L}_1^i + \mathcal{L}_2$$

intermediate supervision on $X_{Q,new}^{i}$



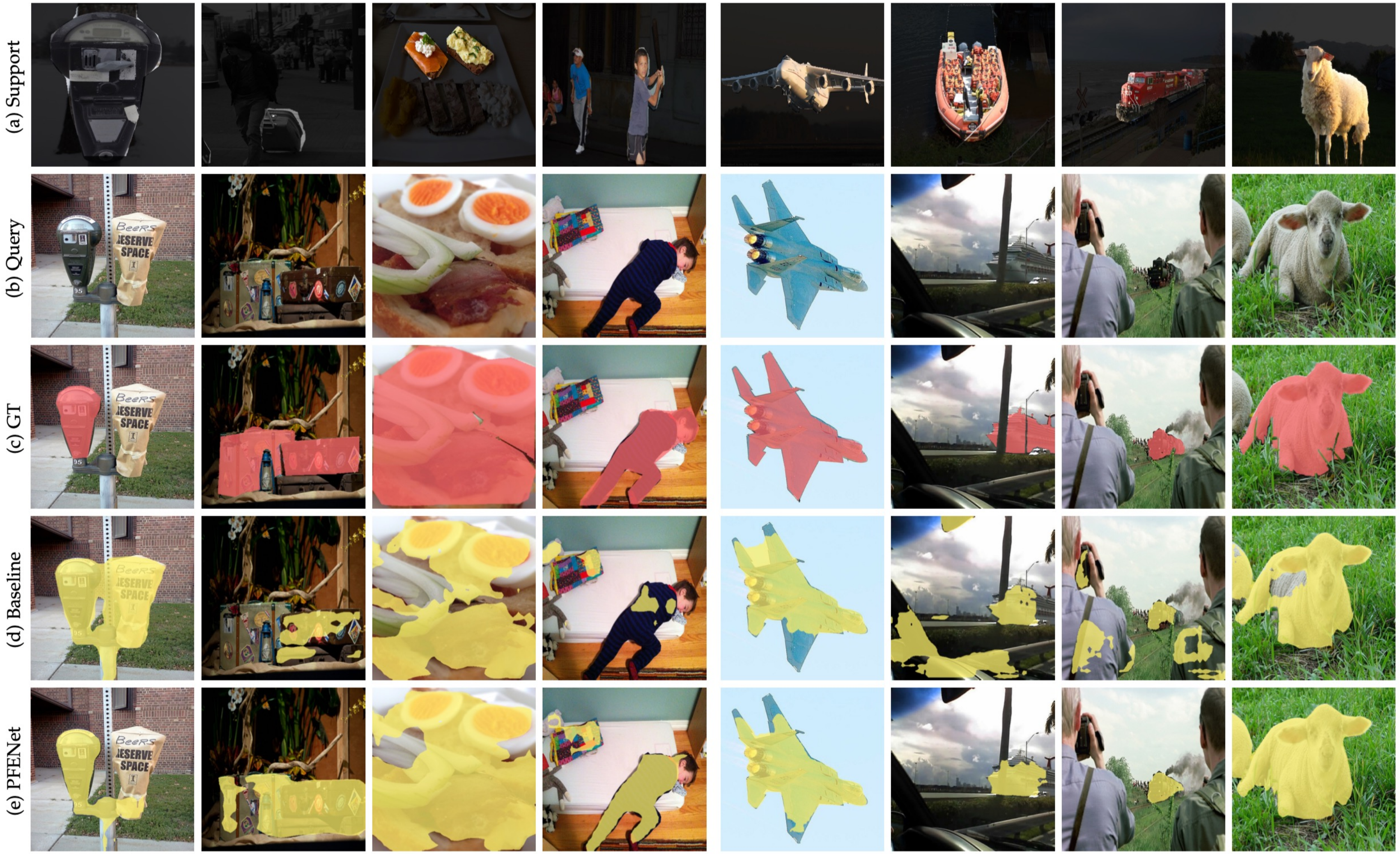
Related Work – PFENet – Inter-Scale Interaction

$$X_{Q,new}^{i} = \mathcal{M}(X_{Q,m}^{Main,i}, X_{Q,m}^{Aux,i})$$



- **Top-down path** adaptively passing information from finer features to the coarse ones
- M that interacts between different scales by selectively **passing useful information**
- α is a 1×1 convolution, β is two 3×3 convolutions

Related Work – PFENet – Result



- Baseline only has one scale size in FEM, not multi-scaled

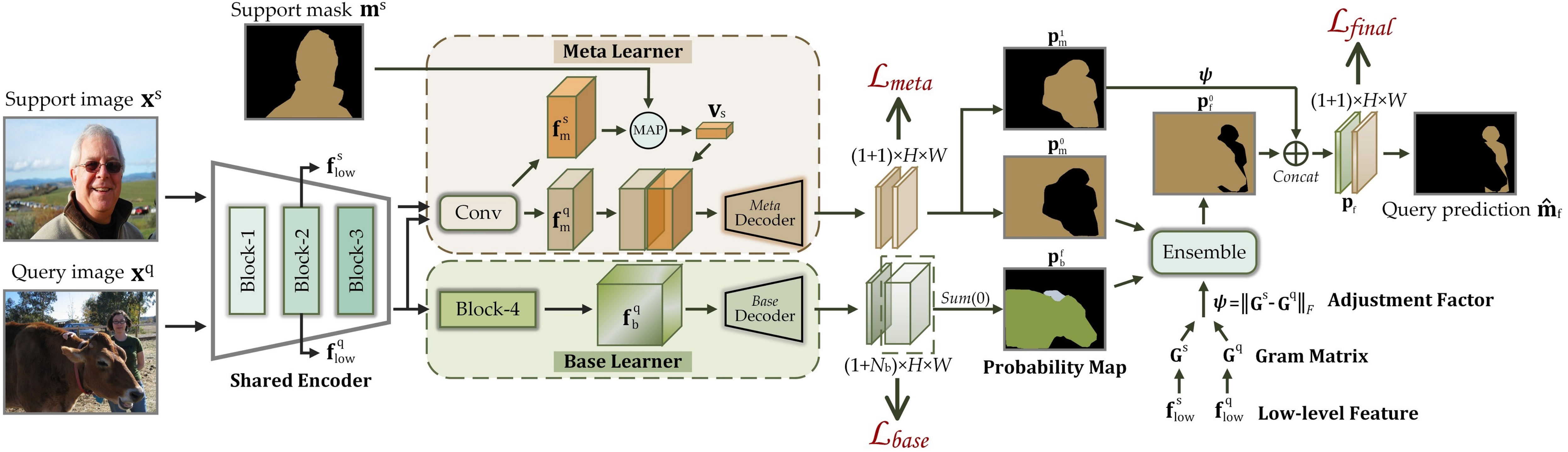
Related Work – PFENet – Contribution

- presented PFENet with the proposed **prior generation** method and the **feature enrichment module (FEM)**
- The prior generation method leverages the cosine-similarity on pre-trained high-level features. Encourages the model to **localize the query targe**.
- FEM helps **solve the spatial inconsistency** by adaptively merging the query and support features at multiple scales with intermediate supervision and conditioned feature selection.

Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

Framework

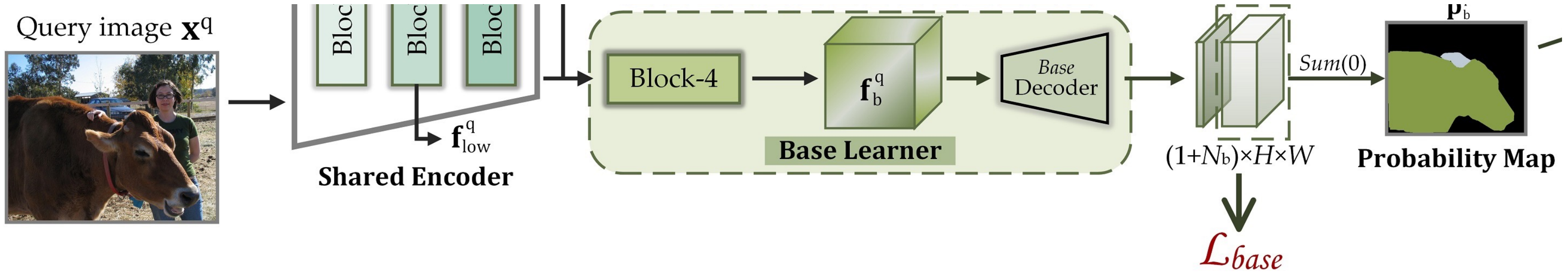


Support mask $\mathbf{m}^s$
Support image $\mathbf{x}^s$
Query image $\mathbf{x}^q$
Block-1
Block-2
Block-3
$\mathbf{f}^s_{low}$
$\mathbf{f}^q_{low}$
Shared Encoder
Meta Learner
Conv
$\mathbf{f}^s_m$
MAP
$\mathbf{v}_s$
$\mathbf{f}^q_m$
Meta Decoder
Block-4
$\mathbf{f}^q_b$
Base Decoder
Base Learner
$\mathcal{L}_{meta}$
(1+1)×H×W
(1+N_b)×H×W
$\mathcal{L}_{base}$
Sum(0)
$\mathbf{p}^1_m$
$\mathbf{p}^0_m$
$\mathbf{p}^f_b$
Probability Map
ψ
$\mathbf{p}^0_f$
Ensemble
Concat
$\mathcal{L}_{final}$
(1+1)×H×W
$\mathbf{p}_f$
Query prediction $\hat{\mathbf{m}}_f$
$\psi = \|\mathbf{G}^s - \mathbf{G}^q\|_F$ Adjustment Factor
$\mathbf{G}^s$ $\mathbf{G}^q$ Gram Matrix
$\mathbf{f}^s_{low}$ $\mathbf{f}^q_{low}$ Low-level Feature

Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

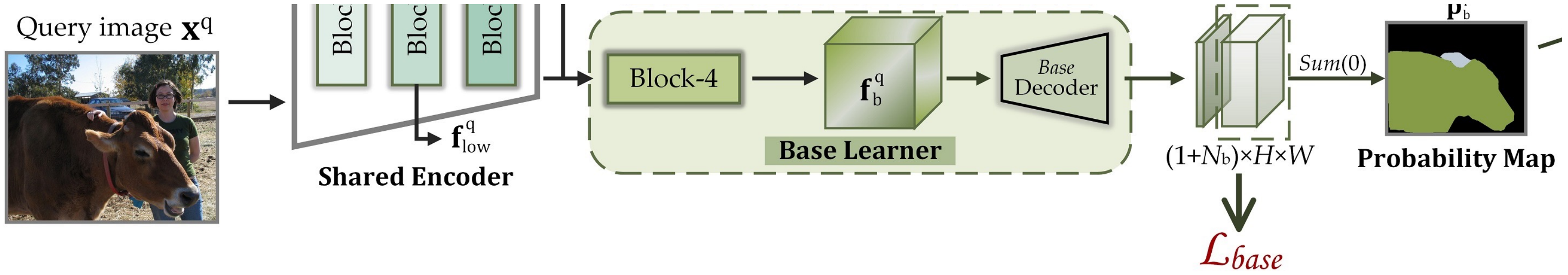
Methods – Base Learner



- Current FSS models are biased towards the seen classes, which impedes the recognition of novel concepts

- Add an additional branch, the base learner, to explicitly predict the regions of base classes in the query images

Methods – Base Learner

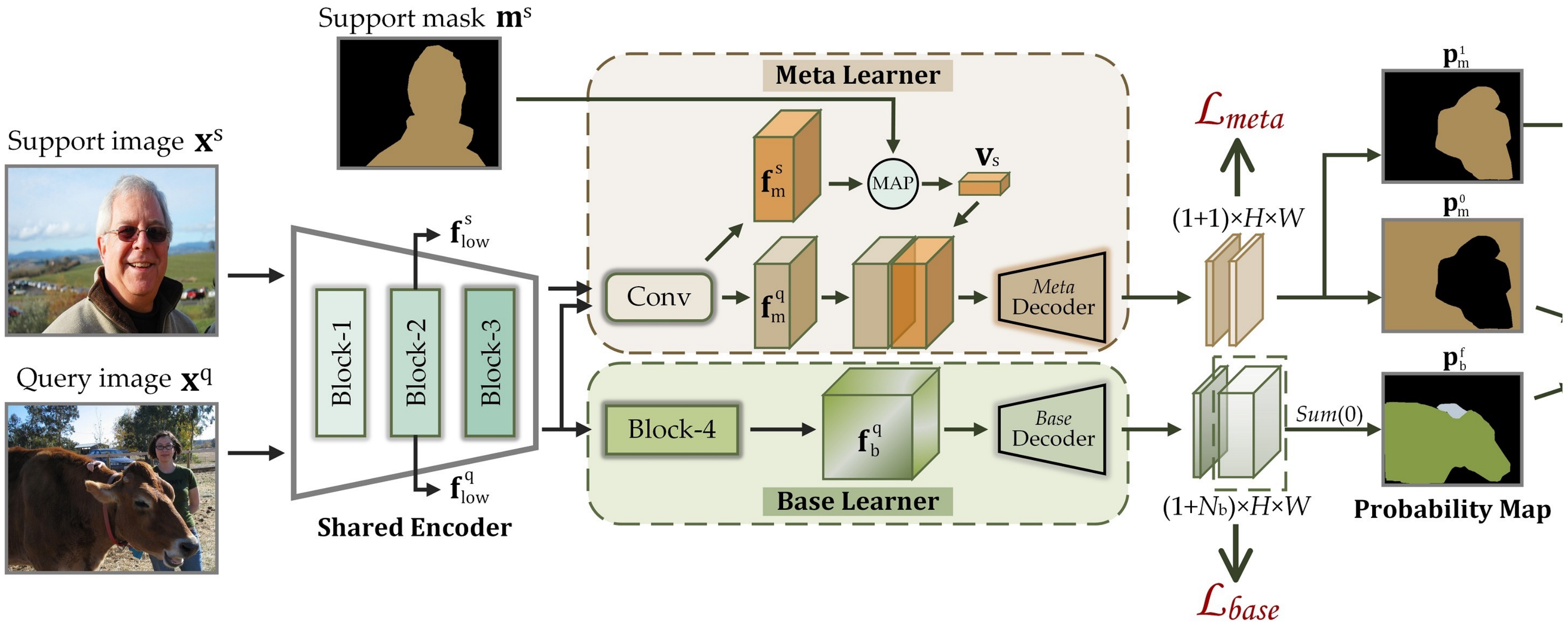


$$\mathbf{f}_{\mathrm{b}}^{\mathrm{q}} = \mathcal{F}_{\mathrm{conv}}\left(\mathcal{E}\left(\mathbf{x}^{\mathrm{q}}\right)\right) \in \mathbb{R}^{c\times h\times w}$$

$$\mathbf{p}_{\mathrm{b}} = \mathrm{softmax}\left(\mathcal{D}_{\mathrm{b}}\left(\mathbf{f}_{\mathrm{b}}^{\mathrm{q}}\right)\right) \in \mathbb{R}^{(1+N_{\mathrm{b}})\times H\times W}$$

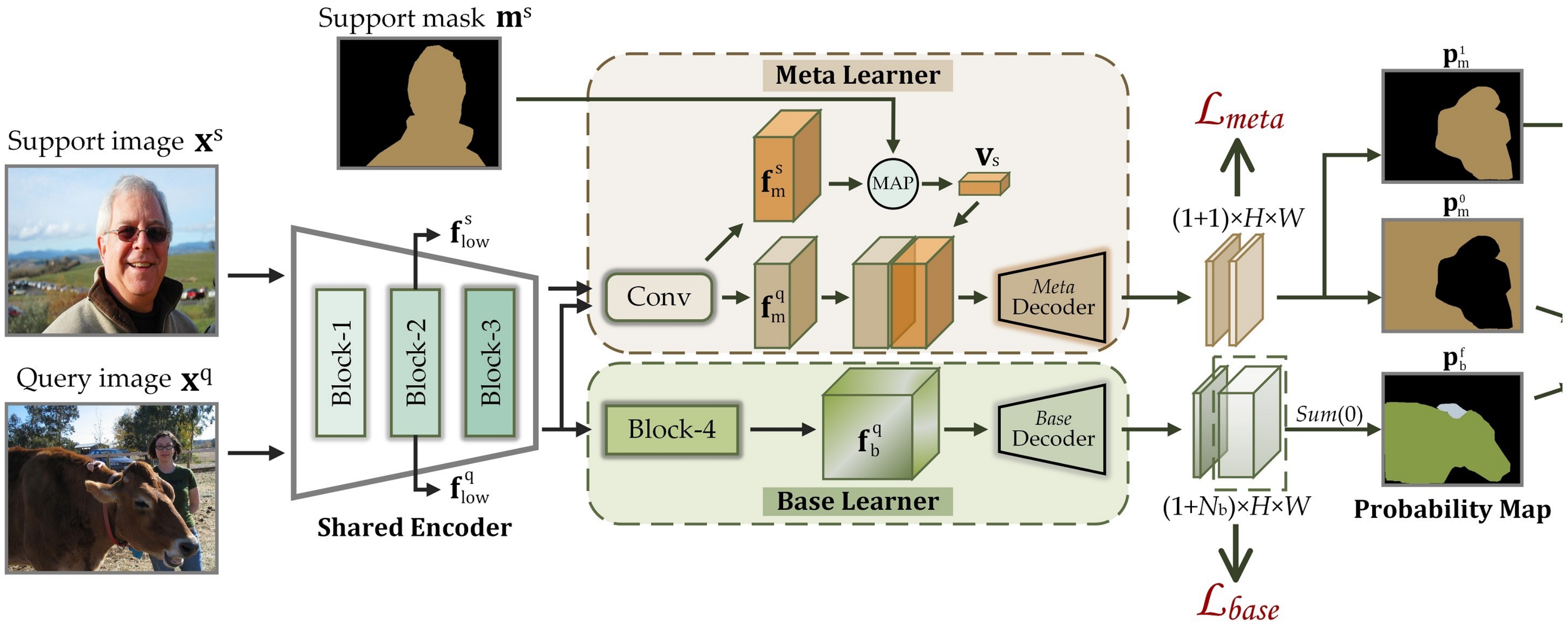
$$\mathcal{L}_{\mathrm{base}} = \frac{1}{n_{\mathrm{bs}}}\sum_{i=1}^{n_{\mathrm{bs}}}\mathrm{CE}\left(\mathbf{p}_{\mathrm{b};i}, \mathbf{m}_{\mathrm{b};i}^{\mathrm{q}}\right)$$

Methods – Meta Learner



- Given a support set $S = \{x^s, m^s\}$ and a query image x^q, the goal of the meta learner is to segment the objects in x^q that share the same category as the annotation mask m^s under the guidance of S.

Methods – Meta Learner



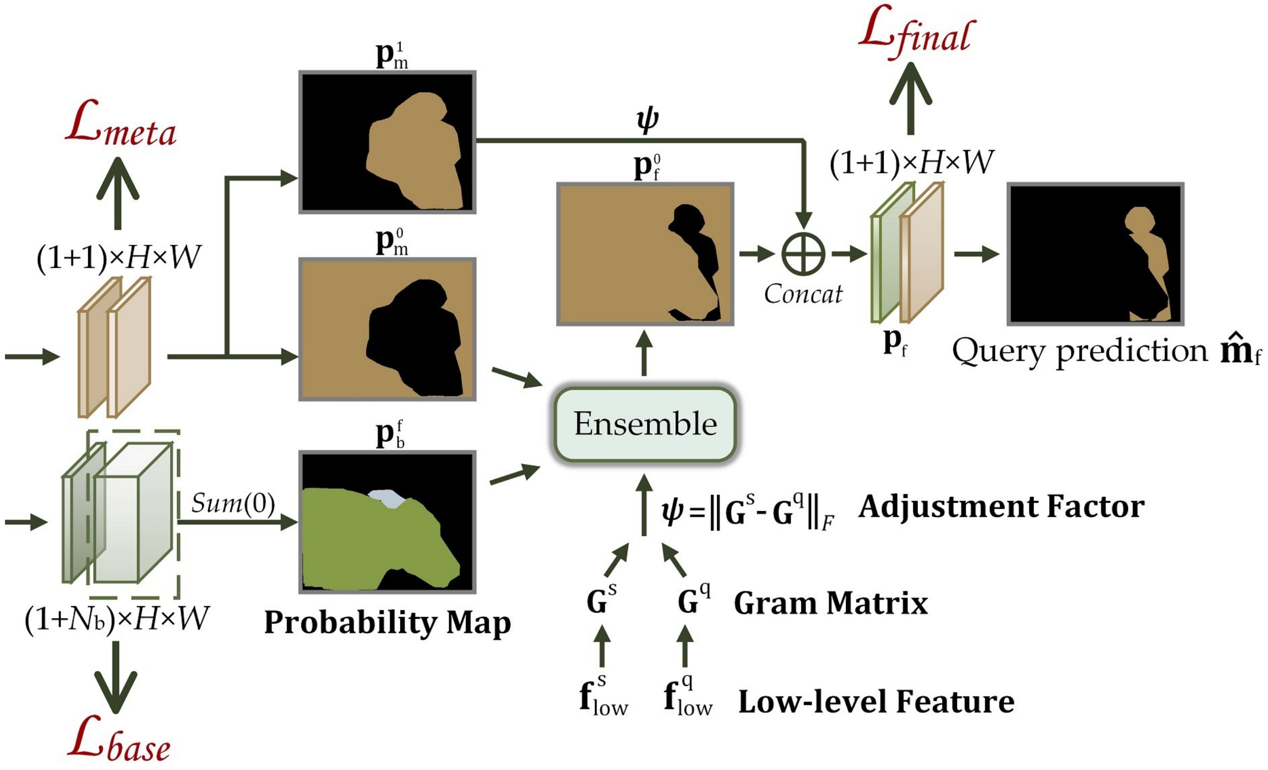
$$\mathbf{f}^s_m = \mathcal{F}_{1\times1}\left(\mathcal{E}\left(\mathbf{x}^s\right)\right) \in \mathbb{R}^{c\times h\times w}$$

$$\mathbf{f}^q_m = \mathcal{F}_{1\times1}\left(\mathcal{E}\left(\mathbf{x}^q\right)\right) \in \mathbb{R}^{c\times h\times w}$$

$$\mathbf{v}_s = \mathcal{F}_{\mathrm{pool}}\left(\mathbf{f}^s_m \odot \mathcal{I}\left(\mathbf{m}^s\right)\right) \in \mathbb{R}^{c}$$

$$\mathbf{p}_m = \mathrm{softmax}\left(\mathcal{D}_m\left(\mathcal{F}_{\mathrm{guidance}}\left(\mathbf{v}_s, \mathbf{f}^q_m\right)\right)\right) \in \mathbb{R}^{2\times H\times W}$$

Methods – Ensemble

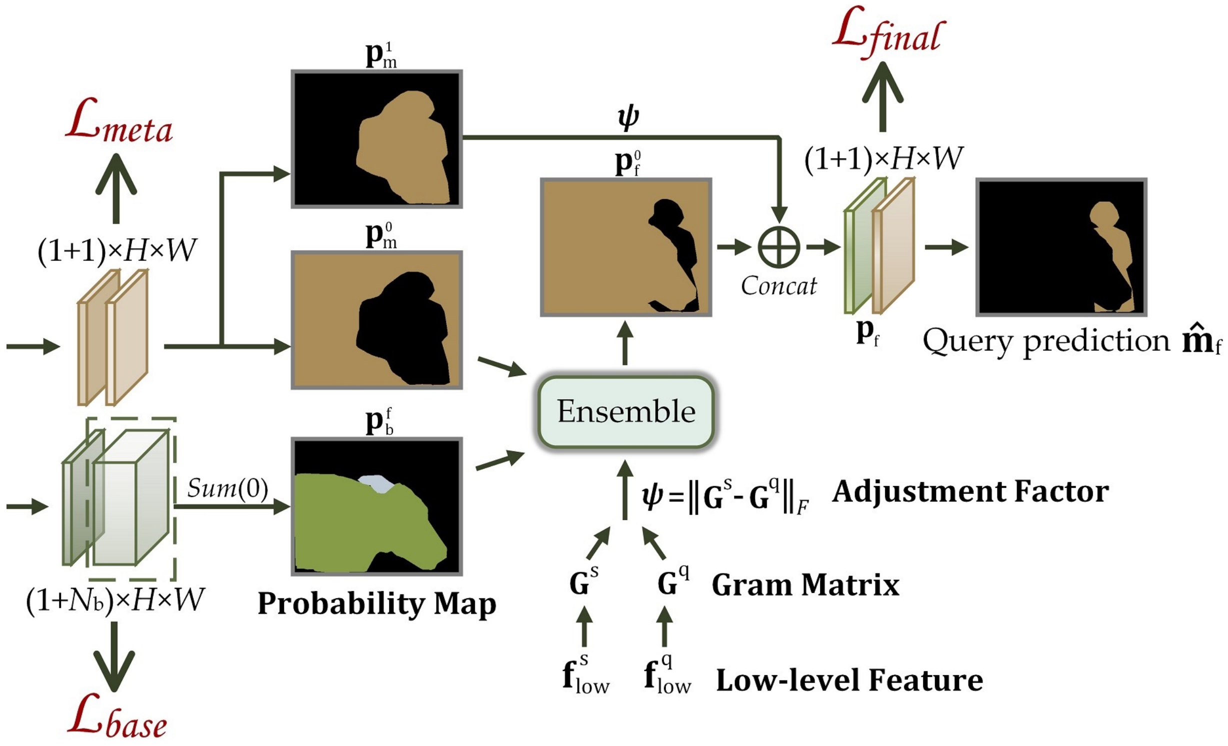


- Meta learners are typically sensitive to the quality of support images, we further propose to leverage the differences between query-support image pairs to **adjust the coarse predictions derived from meta learners**.

- **Integrate the probability maps generated by the base learner** to obtain the prediction of the background region relative to the few-shot task

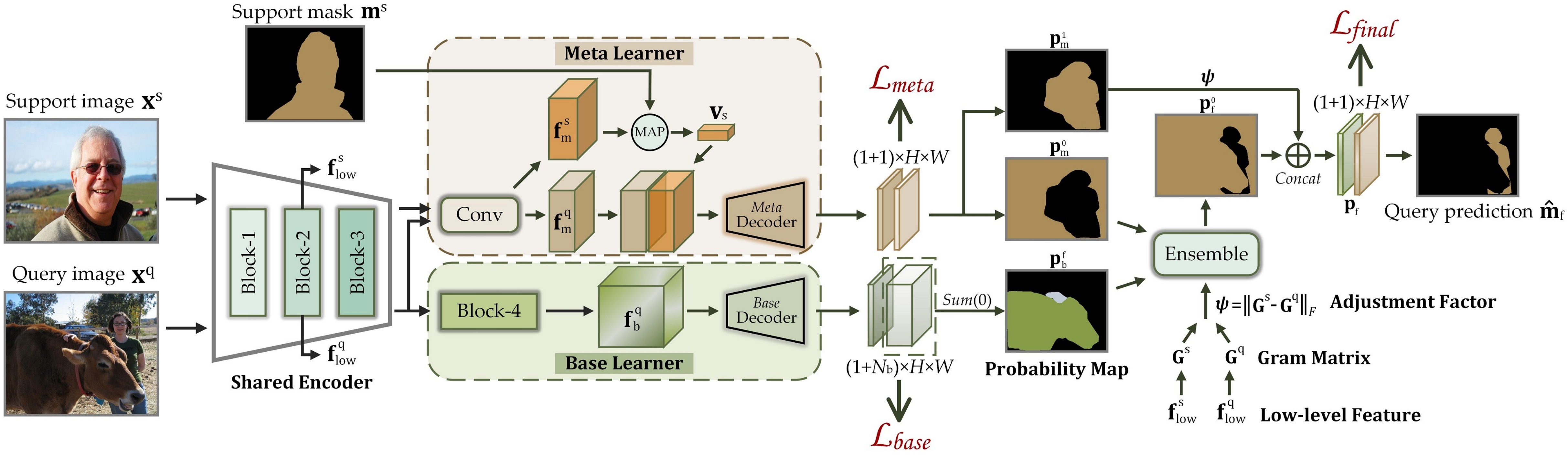
Methods – Ensemble

$$\mathbf{p}_{\mathrm{b}}^{\mathrm{f}} = \sum_{i=1}^{N_{\mathrm{b}}} \mathbf{p}_{\mathrm{b}}^{i}$$



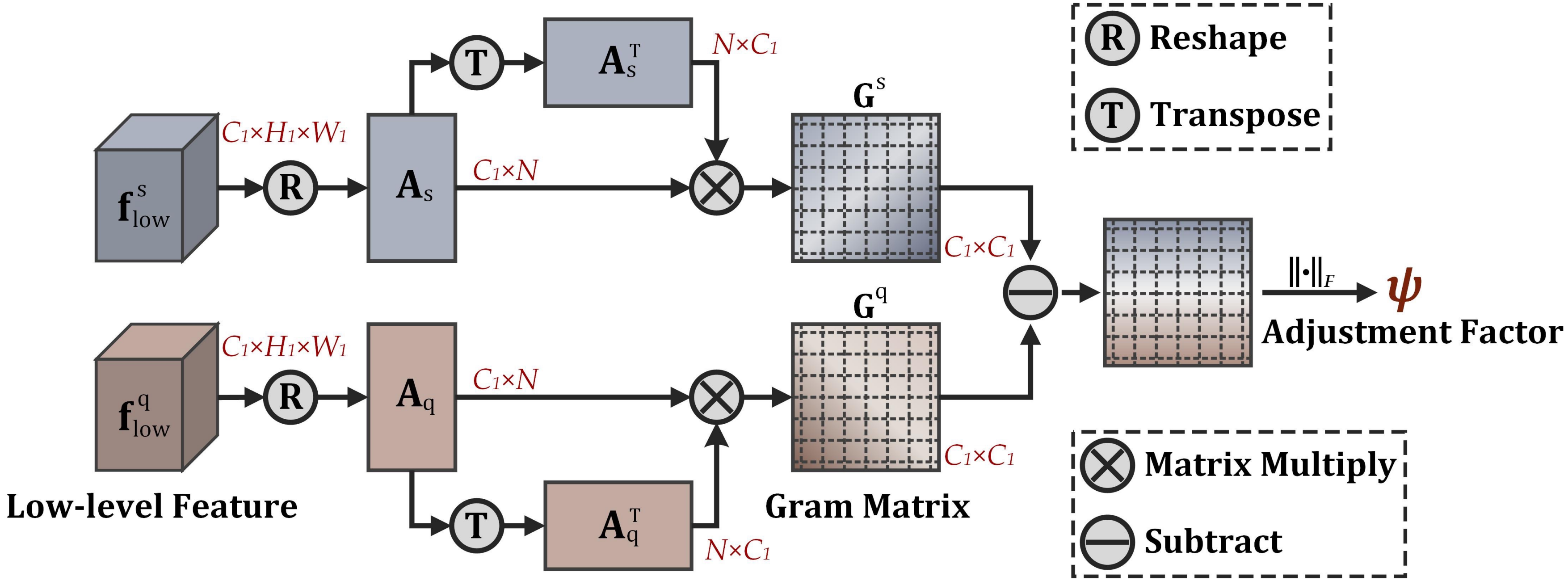
- the superscript of f stands for the foreground, and the subscript b stands for the base learner
- N_b represents the number of base categories

Methods – Ensemble



- we leverage the low-level features $f_{low}^s, f_{low}^q \in \mathbb{R}^{c_1 * H_1 * W_1}$ extracted from the fixed backbone network

Methods – Ensemble

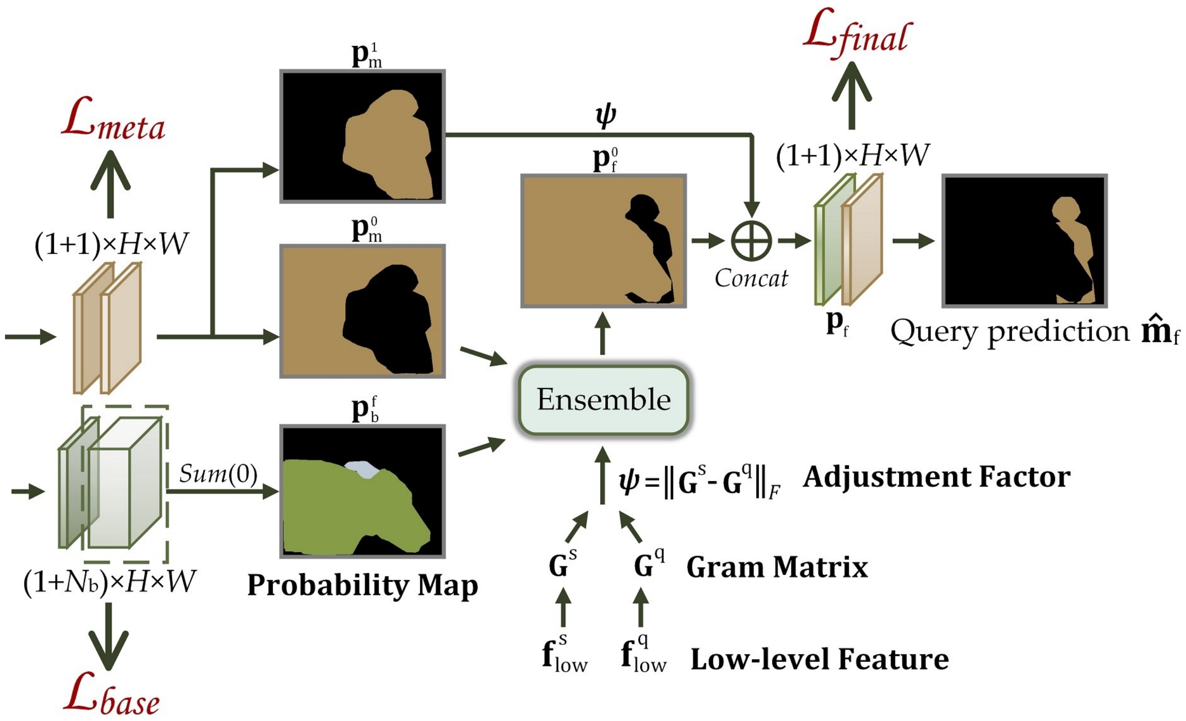


$$\mathbf{A}_{\mathrm{s}} = \mathcal{F}_{\mathrm{reshape}}\left(\mathbf{f}_{\mathrm{low}}^{\mathrm{s}}\right) \in \mathbb{R}^{C_1 \times N},$$

$$\mathbf{G}^{\mathrm{s}} = \mathbf{A}_{\mathrm{s}} \mathbf{A}_{\mathrm{s}}^{\mathsf{T}} \in \mathbb{R}^{C_1 \times C_1},$$

$$\psi = \left\| \mathbf{G}^{\mathrm{s}} - \mathbf{G}^{\mathrm{q}} \right\|_F$$

Methods – Ensemble



$$\mathbf{p}_{\mathrm{f}}^{0}=\mathcal{F}_{\text{ensemble}}\left(\mathcal{F}_{\psi}\left(\mathbf{p}_{\mathrm{m}}^{0}\right), \mathbf{p}_{\mathrm{b}}^{\mathrm{f}}\right),$$

$$\mathbf{p}_{\mathrm{f}}=\mathbf{p}_{\mathrm{f}}^{0} \oplus \mathcal{F}_{\psi}\left(\mathbf{p}_{\mathrm{m}}^{1}\right),$$

$$\mathcal{L}=\mathcal{L}_{\text{final}}+\lambda \mathcal{L}_{\text{meta}},$$

$$\mathcal{L}_{\text{final}}=\frac{1}{n_{\mathrm{e}}} \sum_{i=1}^{n_{\mathrm{e}}} \operatorname{BCE}\left(\mathbf{p}_{i}^{\mathrm{q}}, \mathbf{m}_{i}^{\mathrm{q}}\right),$$

Methods – K-Shot setting

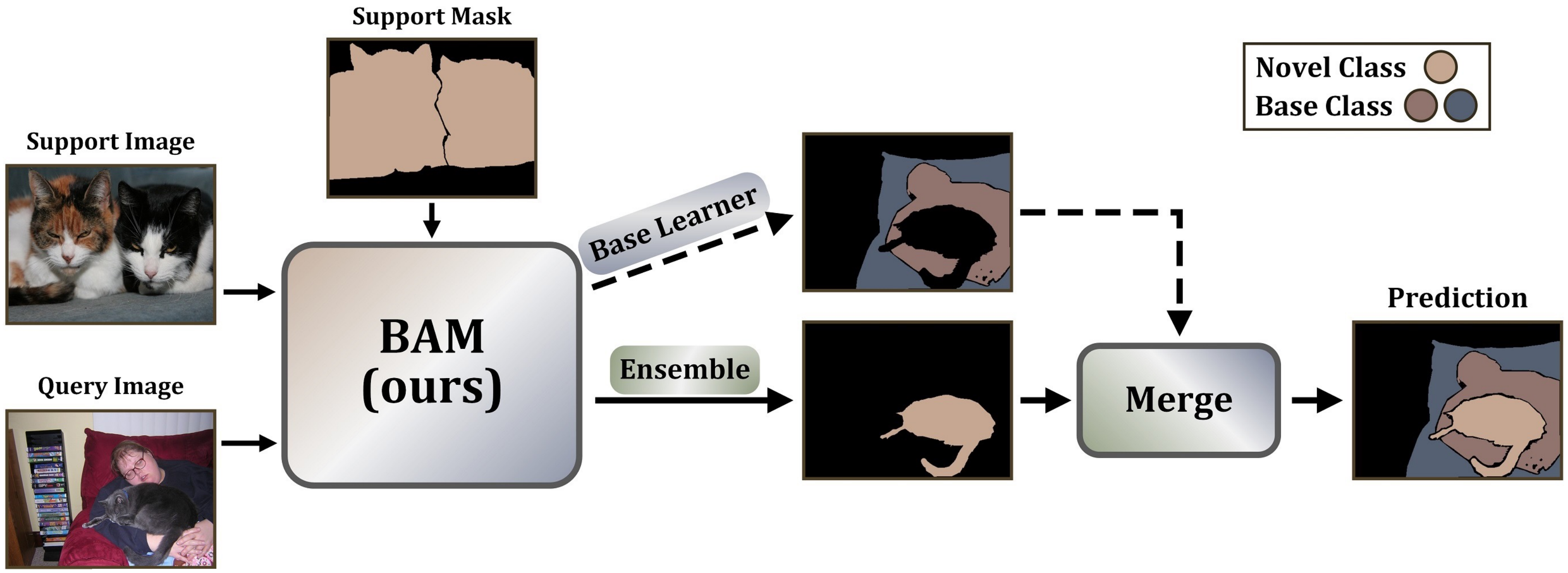
- the task is extended to the K-shot (K>1) setting, more than one annotated (support) images
- Two **FC layers** are applied to adaptively estimate the weight of each support image based on the adjustment factor ψ

$$\psi_{\mathrm{t}} \in \mathbb{R}^{K}$$

$$\eta = \operatorname{soft\,max}\left(\mathbf{w}_2^{\mathsf{T}} \mathrm{ReLU}\left(\mathbf{w}_1^{\mathsf{T}} \psi_{\mathrm{t}}\right)\right) \in \mathbb{R}^{K}$$

Methods – Extension to Generalized FSS

(c)



$$\hat{\mathbf{m}}_{\mathrm{g}}^{(x,y)} = \begin{cases} 1 & \mathbf{p}_{\mathrm{f}}^{1;(x,y)} > \tau \\ \hat{\mathbf{m}}_{\mathrm{b}}^{(x,y)} & \mathbf{p}_{\mathrm{f}}^{1;(x,y)} \leq \tau \text{ and } \hat{\mathbf{m}}_{\mathrm{b}}^{(x,y)} \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{\mathbf{m}}_{\mathrm{b}} = \arg\max\left(\mathbf{p}_{\mathrm{b}}\right) \in \{0, 1, \dots, N_{\mathrm{b}}\}^{H \times W}$$

Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

Experiments – Setup

- Evaluate the performance on two FSS datasets, namely **PASCAL-5i** and **COCO-20i**
- Divided into **four folds**, and the experiments are conducted in a cross validation manner. For each fold, we randomly sample **1,000 pairs** of support and query images for validation.
- Divided into two stages, **pretraining** and **meta-training**
- Standard **supervised learning** to train **the base learner**
- Jointly train **the meta learner** and **ensemble** module in an **episodic learning fashion**, and the parameters of the base learner are fixed in this stage.
- Two learners share the **same encoder** to extract the features of input images, which is also **not optimized** to facilitate generalization.

Experiments – Result

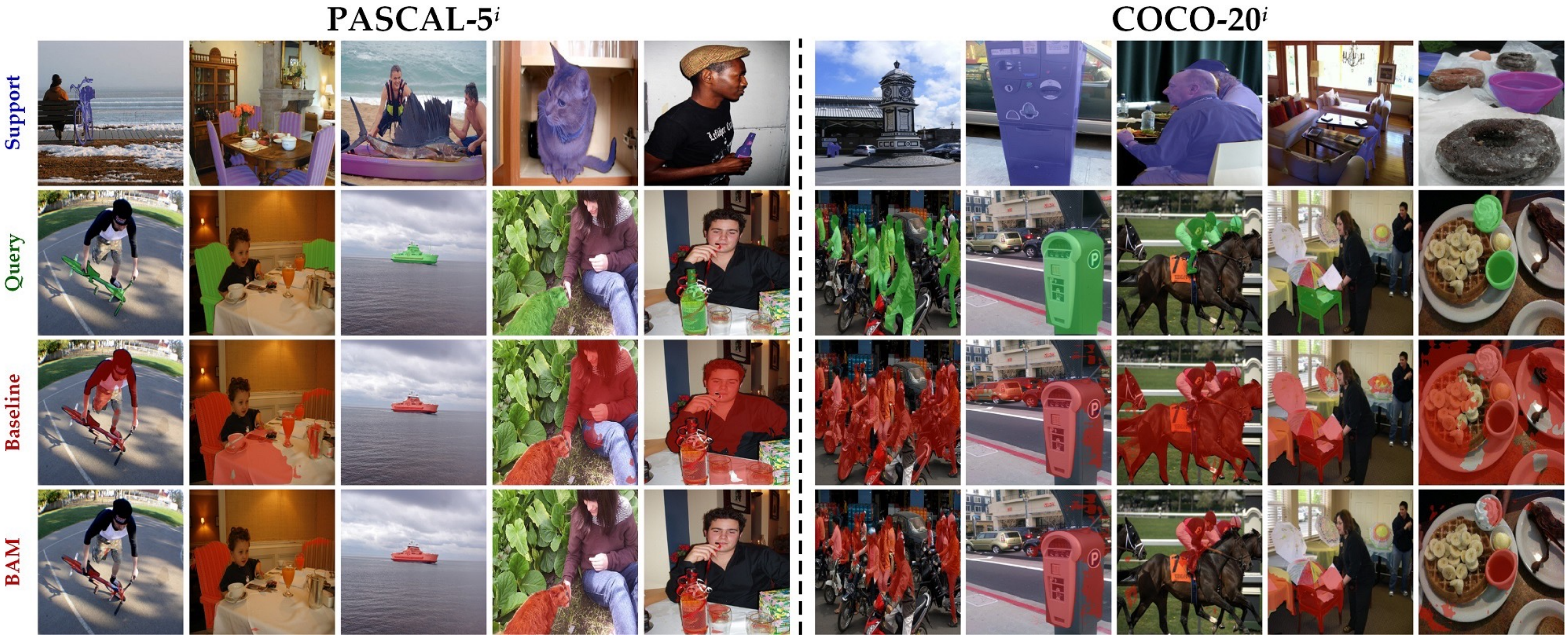
COCO-20i

Backbone	Method	1-shot					5-shot				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
VGG16	FWB [38]	18.35	16.72	19.59	25.43	20.02	20.94	19.24	21.94	28.39	22.63
	PFENet [51]	35.40	38.10	36.80	34.70	36.30	38.20	42.50	41.80	38.90	40.40
	PRNet [32]	27.46	32.99	26.70	28.98	29.03	31.18	36.54	31.54	32.00	32.82
	Baseline	38.42	43.75	44.32	39.84	41.58	45.93	48.88	47.87	46.96	47.41
	BAM (ours)	**38.96**	**47.04**	**46.41**	**41.57**	**43.50**	**47.02**	**52.62**	**48.59**	**49.11**	**49.34**
ResNet50	HFA [31]	28.65	36.02	30.16	33.28	32.03	32.69	42.12	30.35	36.19	35.34
	ASGNet [23]	-	-	-	-	34.56	-	-	-	-	42.48
	HSNet [37]	36.30	43.10	38.70	38.70	39.20	43.30	51.30	48.20	45.00	46.90
	Baseline	41.92	45.35	43.86	41.24	43.09	46.98	51.87	49.49	47.81	49.04
	BAM (ours)	**43.41**	**50.59**	**47.49**	**43.42**	**46.23**	**49.26**	**54.20**	**51.63**	**49.55**	**51.16**

PASCAL-5i

Backbone	Method	1-shot					5-shot				
		Fold-0	Fold-1	Fold-2	Fold-3	Mean	Fold-0	Fold-1	Fold-2	Fold-3	Mean
VGG16	SG-One (TCYB'19) [67]	40.20	58.40	48.40	38.40	46.30	41.90	58.60	48.60	39.40	47.10
	PANet (ICCV'19) [56]	42.30	58.00	51.10	41.20	48.10	51.80	64.60	59.80	46.50	55.70
	FWB (ICCV'19) [56]	47.00	59.60	52.60	48.30	51.90	50.90	62.90	56.50	50.10	55.10
	CRNet (CVPR'20) [33]	-	-	-	-	55.20	-	-	-	-	58.50
	PFENet (TPAMI'20) [51]	56.90	68.20	54.40	52.40	58.00	59.00	69.10	54.80	52.90	59.00
	HSNet (ICCV'21) [37]	59.60	65.70	59.60	54.00	59.70	64.90	69.00	64.10	58.60	64.10
	Baseline	59.90	67.51	64.93	55.72	62.02	64.02	71.51	69.39	63.55	67.12
	BAM (ours)	**63.18**	**70.77**	**66.14**	**57.53**	**64.41**	**67.36**	**73.05**	**70.61**	**64.00**	**68.76**
ResNet50	CANet (ICCV'19) [66]	52.50	65.90	51.30	51.90	55.40	55.50	67.80	51.90	53.20	57.10
	PGNet (ICCV'19) [65]	56.00	66.90	50.60	50.40	56.00	57.70	68.70	52.90	54.60	58.50
	CRNet (CVPR'20) [33]	-	-	-	-	55.70	-	-	-	-	58.80
	PPNet (ECCV'20) [34]	48.58	60.58	55.71	46.47	52.84	58.85	68.28	66.77	57.98	62.97
	PFENet (TPAMI'20) [51]	61.70	69.50	55.40	56.30	60.80	63.10	70.70	55.80	57.90	61.90
	HSNet (ICCV'21) [37]	64.30	70.70	60.30	60.50	64.00	70.30	73.20	67.40	67.10	69.50
	Baseline	65.68	71.41	65.56	58.93	65.40	67.28	72.38	69.16	66.25	68.77
	BAM (ours)	**68.97**	**73.59**	**67.55**	**61.13**	**67.81**	**70.59**	**75.05**	**70.79**	**67.20**	**70.91**

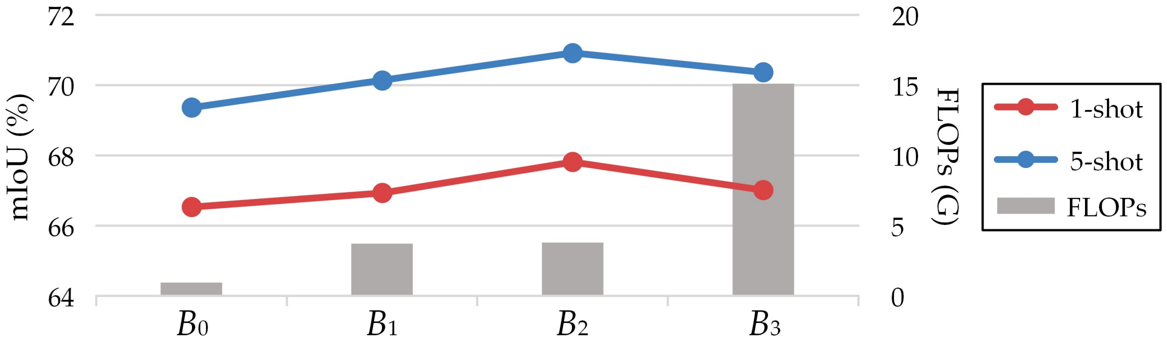
Experiments – Comparison



Experiments – Ablation Study

- Two learners could be **trained jointly or separately**. In our experiments, the latter scheme exhibits better performance. Since base learner tends to **fix the parameters to enhance generalization**, while meta-learner tends to **update the parameters to extract more discriminative features**

- low-level features f_{low} to estimate ψ with ResNet50 backbone. B_i denotes the feature maps extracted from the i-th convolutional blocks. **B_2 features shows a better trade-off** between segmentation accuracy and computational complexity.



Outline

- Introduction
- Related Work
- Framework
- Method
- Experiment
- Conclusion

Conclusion

- Proposed a novel scheme **to alleviate the bias problem** of FSS models towards the seen concepts.
- The core idea of our scheme is to leverage the **base learner to identify the confusable (base) regions** in the query images and further refine the prediction of the meta learner.
- Moreover, extended the current task to a **more challenging generalized** setting and produced strong baseline results.