Weakly-supervised Object Representation Learning for Few-shot Semantic Segmentation

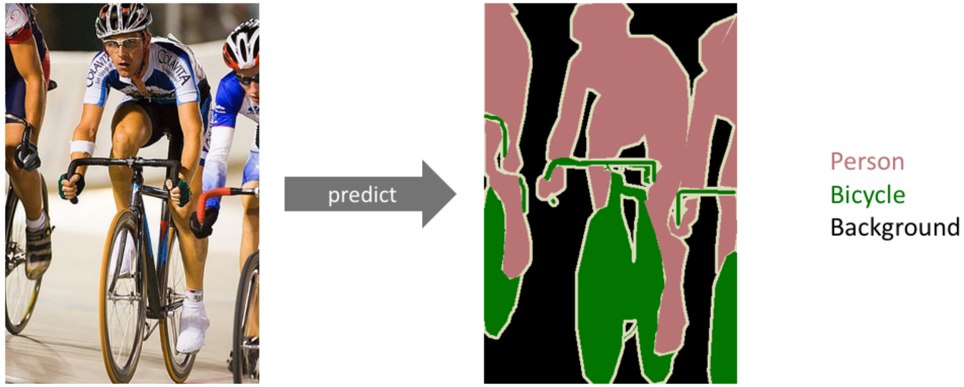
WACV 2021

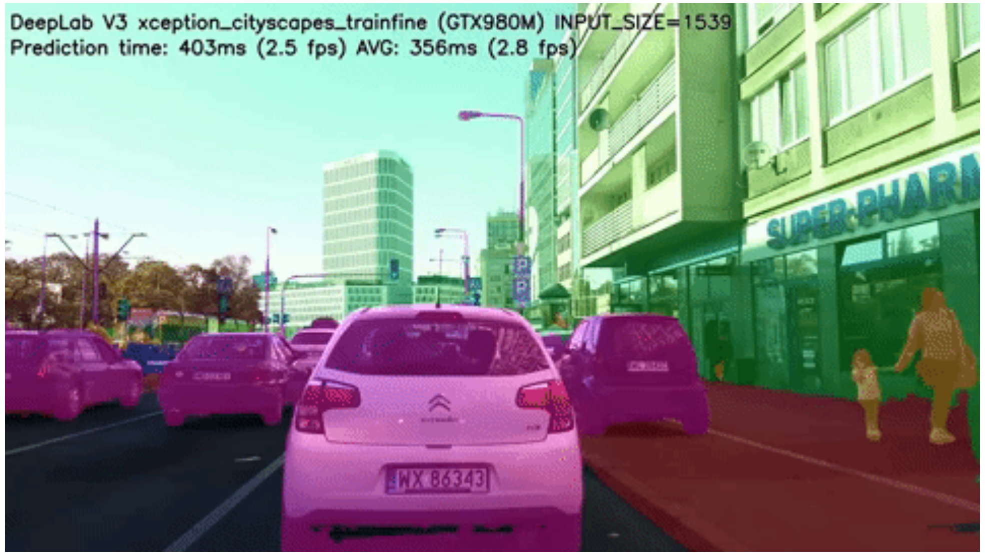
Xiaowen Ying, Xin Li, Mooi Choo Chuah
Lehigh University

Semantic Segmentation

- A task of assigning a class label to each pixel in the image.

- One of the fundamental tasks in Computer Vision.




DeepLab V3 xception_cityscapes_trainfine (GTX980M) INPUT_SIZE=1539
Prediction time: 403ms (2.5 fps) AVG: 356ms (2.8 fps)

New challenges in semantic segmentation

Data Labeling is Expensive

Training a semantic segmentation model requires large amount of pixel-wise annotated images, which is costly to obtain.

Limited to Segment Predefined Categories

Once the training is done, the model is limited to segment those predefined classes in training set.

Few-shot Segmentation

Goal: Perform segmentation on unseen categories merely based on one or a few support examples.

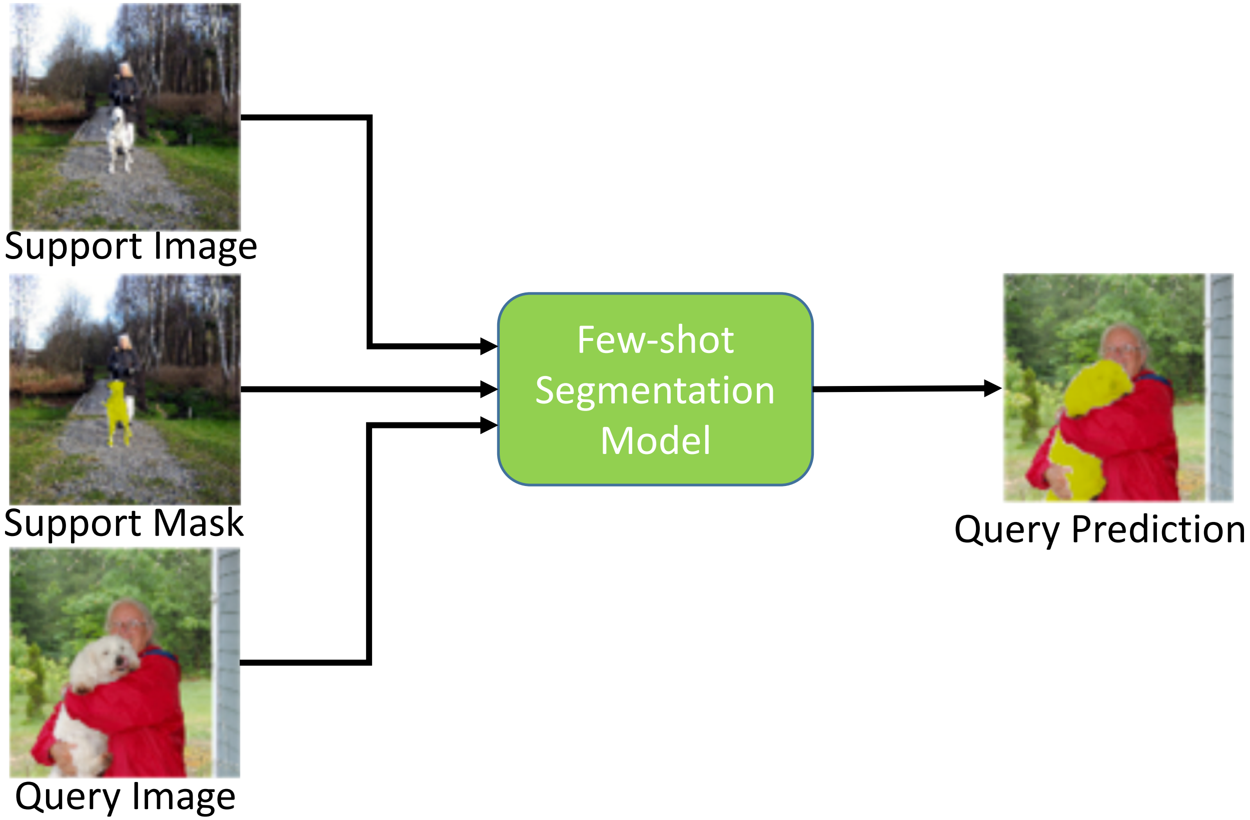


Illustration of one-shot segmentation

Motivation of Our Design

Key to this problem:

Effectively utilizing object information from support examples.

- Existing methods typically generate object-level representations by **averaging foreground features** in support images.

- We found that such object representations are typically **noisy** and **less distinguishable**.

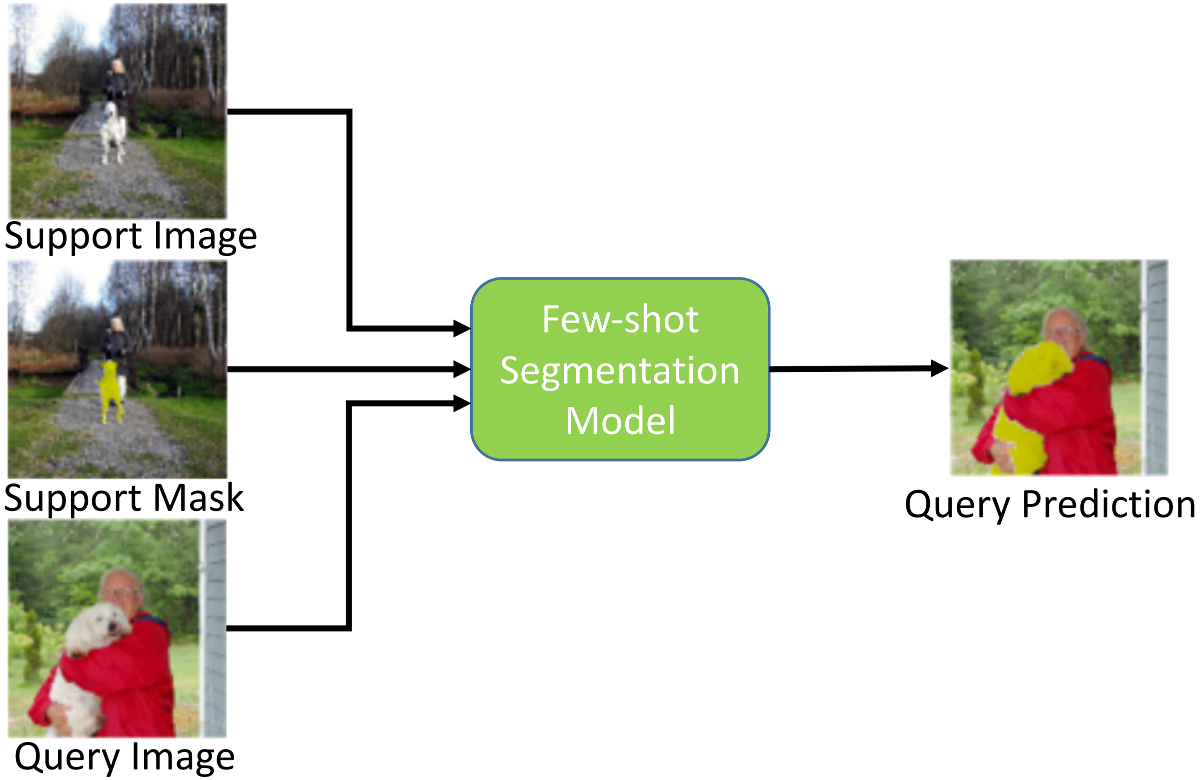


Illustration of one-shot segmentation

Our Contributions

1. A new few-shot segmentation framework.
2. A novel Object Representation Generator (ORG) module.
3. Weakly-supervised training scheme for the ORG module.
4. SOTA performances on two benchmarks.

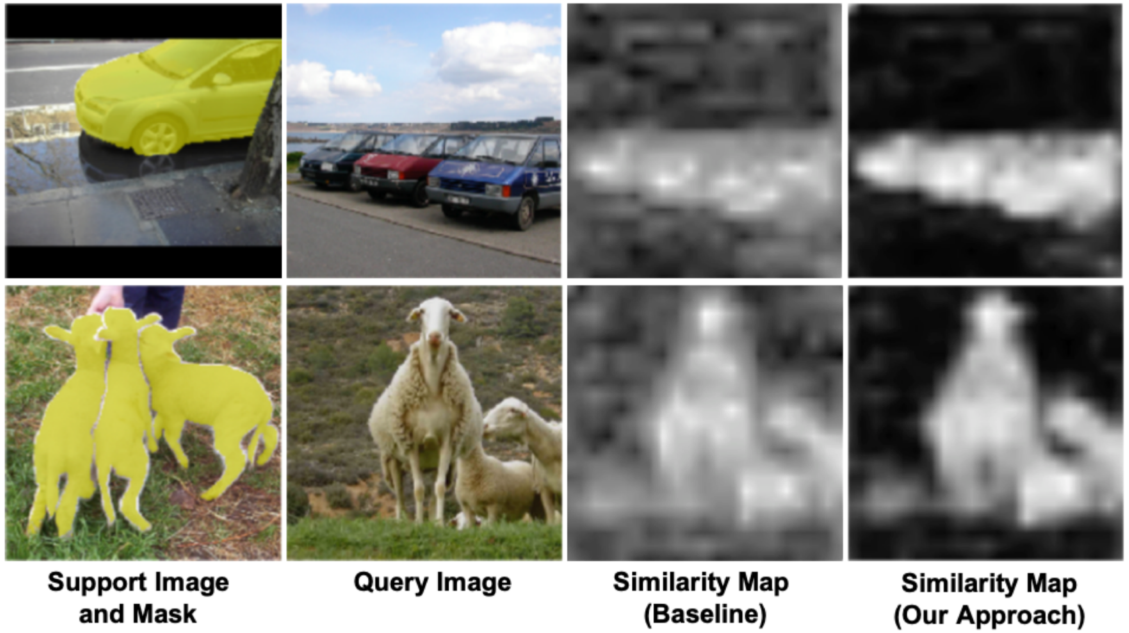
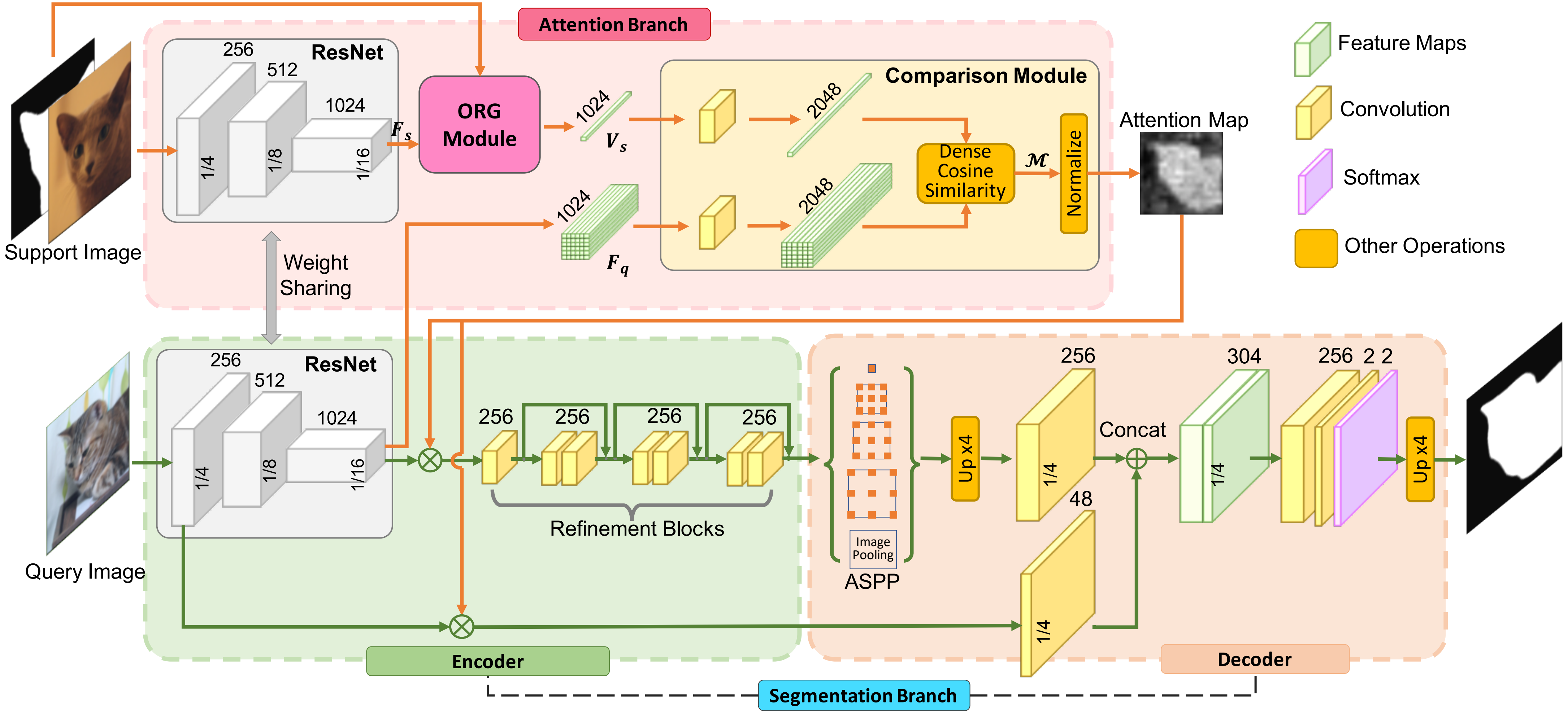


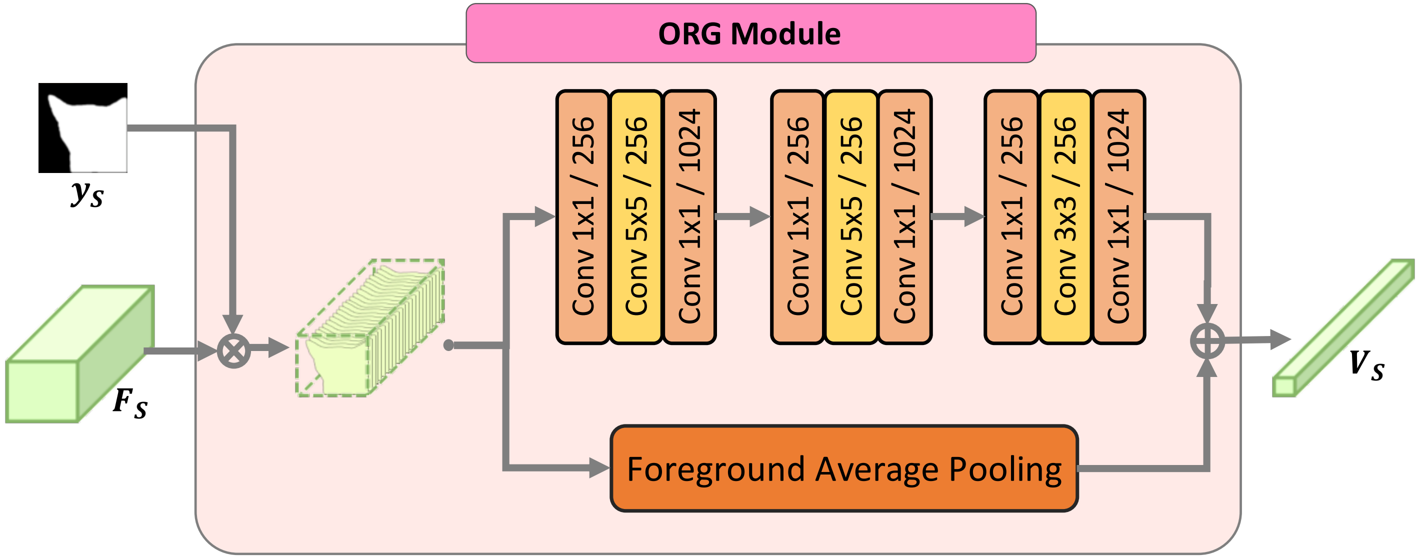
Illustration of the similarity maps produced by different object representation approaches.

Our Architecture



The proposed architecture

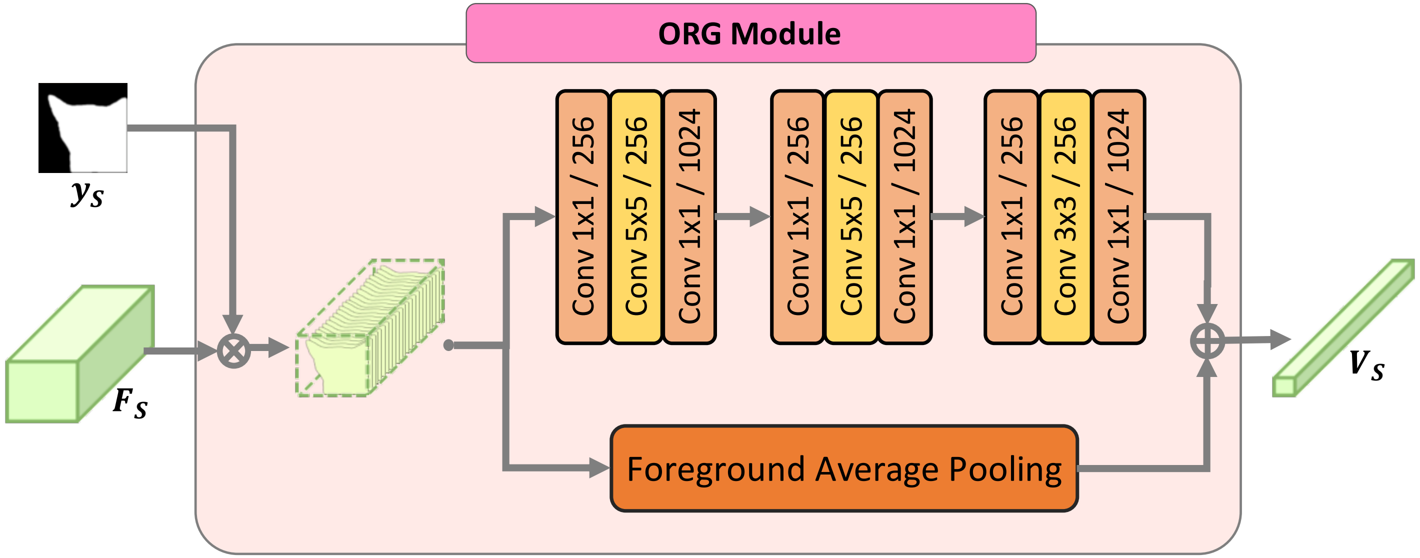
Object Representation Generator



Architecture of the ORG module

- Consist of several convolution blocks that learns to produce object representation.
- Bottleneck block design to reduce number of parameters.
- Add foreground average pooling as a parallel branch.

Object Representation Generator

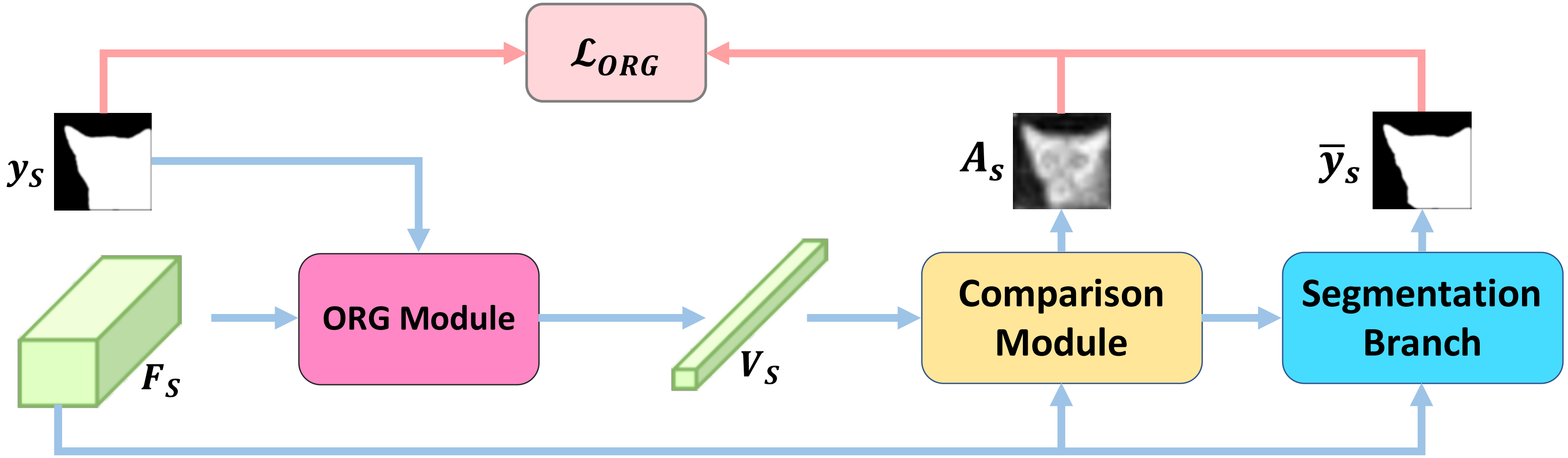


Architecture of the ORG module

- Consist of several convolution blocks that learns to produce object representation.
- Bottleneck block design to reduce number of parameters.
- Add foreground average pooling as a parallel branch.

Problem: How can we teach this module to produce better object representation?

Weakly-supervised Learning for ORG



Computation Graph of the proposed weakly-supervised training scheme

Intuitively, this learning process forces the ORG module to improve the quality of the object representations, such that it can better segment the source object itself in the original support image.

Dataset

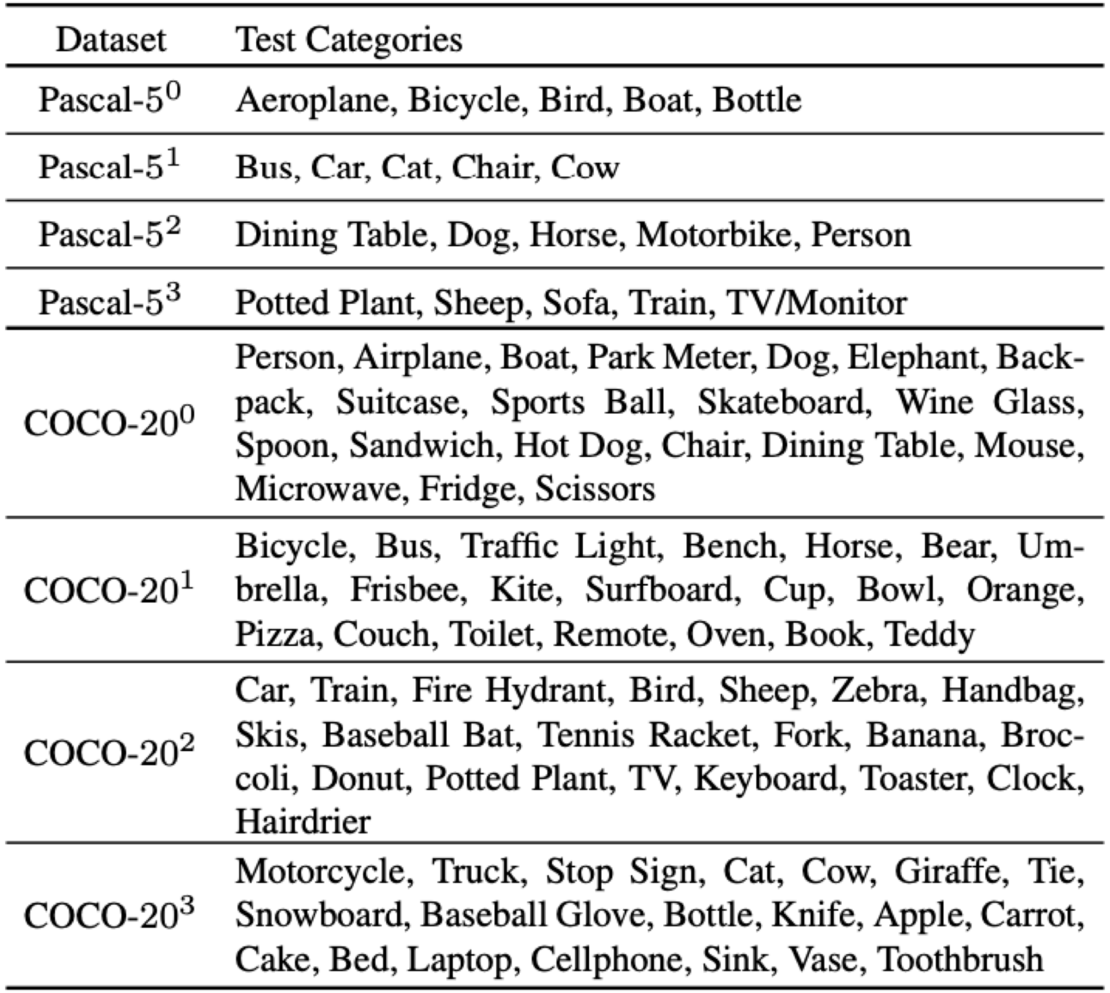
Pascal-5i

- 20 categories in the original PASCAL-VOC dataset are evenly divided into 4 splits for 4-fold cross-validation.
- Each fold consists of 1 split for testing and the other 3 splits for training.

COCO-20i

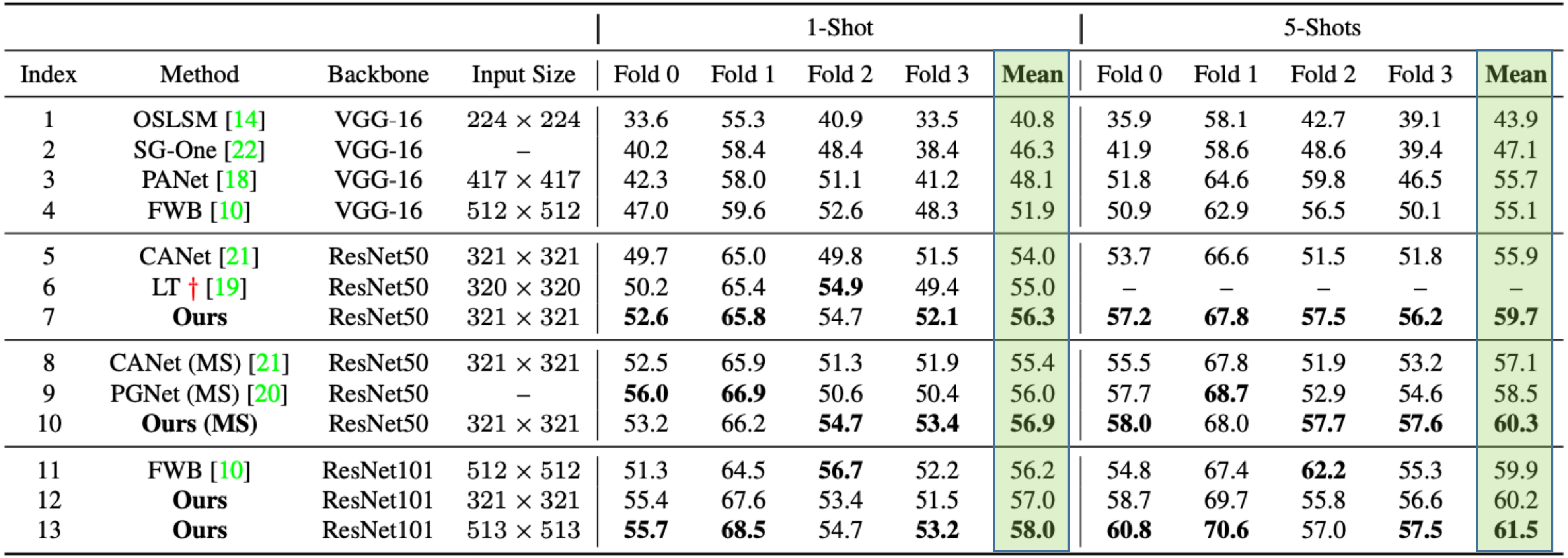
- 80 categories in the original MSCOCO dataset are evenly divided into 4 splits for 4-fold cross-validation.
- Each fold consists of 1 split for testing and the other 3 splits for training.

Summary of testing object categories used in each fold



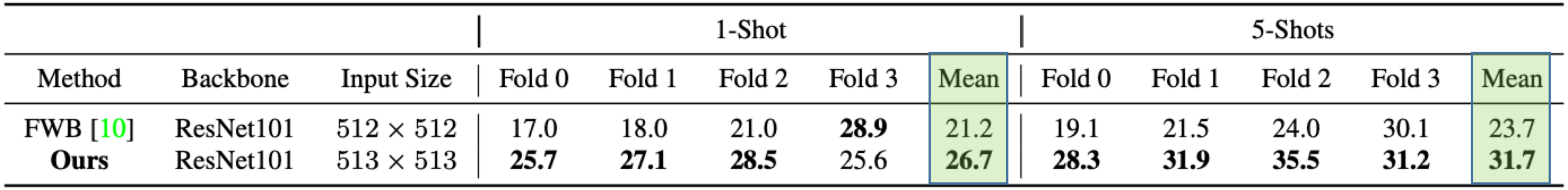
Dataset	Test Categories
Pascal-5^0	Aeroplane, Bicycle, Bird, Boat, Bottle
Pascal-5^1	Bus, Car, Cat, Chair, Cow
Pascal-5^2	Dining Table, Dog, Horse, Motorbike, Person
Pascal-5^3	Potted Plant, Sheep, Sofa, Train, TV/Monitor
COCO-20^0	Person, Airplane, Boat, Park Meter, Dog, Elephant, Backpack, Suitcase, Sports Ball, Skateboard, Wine Glass, Spoon, Sandwich, Hot Dog, Chair, Dining Table, Mouse, Microwave, Fridge, Scissors
COCO-20^1	Bicycle, Bus, Traffic Light, Bench, Horse, Bear, Umbrella, Frisbee, Kite, Surfboard, Cup, Bowl, Orange, Pizza, Couch, Toilet, Remote, Oven, Book, Teddy
COCO-20^2	Car, Train, Fire Hydrant, Bird, Sheep, Zebra, Handbag, Skis, Baseball Bat, Tennis Racket, Fork, Banana, Broccoli, Donut, Potted Plant, TV, Keyboard, Toaster, Clock, Hairdrier
COCO-20^3	Motorcycle, Truck, Stop Sign, Cat, Cow, Giraffe, Tie, Snowboard, Baseball Glove, Bottle, Knife, Apple, Carrot, Cake, Bed, Laptop, Cellphone, Sink, Vase, Toothbrush

Quantitative Results



				1-Shot					5-Shots				
Index	Method	Backbone	Input Size	Fold 0	Fold 1	Fold 2	Fold 3	**Mean**	Fold 0	Fold 1	Fold 2	Fold 3	**Mean**
1	OSLSM [14]	VGG-16	224 × 224	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9
2	SG-One [22]	VGG-16	–	40.2	58.4	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1
3	PANet [18]	VGG-16	417 × 417	42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7
4	FWB [10]	VGG-16	512 × 512	47.0	59.6	52.6	48.3	51.9	50.9	62.9	56.5	50.1	55.1
5	CANet [21]	ResNet50	321 × 321	49.7	65.0	49.8	51.5	54.0	53.7	66.6	51.5	51.8	55.9
6	LT † [19]	ResNet50	320 × 320	50.2	65.4	**54.9**	49.4	55.0	–	–	–	–	–
7	**Ours**	ResNet50	321 × 321	**52.6**	**65.8**	54.7	**52.1**	**56.3**	**57.2**	**67.8**	**57.5**	**56.2**	**59.7**
8	CANet (MS) [21]	ResNet50	321 × 321	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1
9	PGNet (MS) [20]	ResNet50	–	**56.0**	**66.9**	50.6	50.4	56.0	57.7	**68.7**	52.9	54.6	58.5
10	**Ours (MS)**	ResNet50	321 × 321	53.2	66.2	**54.7**	**53.4**	**56.9**	**58.0**	68.0	**57.7**	**57.6**	**60.3**
11	FWB [10]	ResNet101	512 × 512	51.3	64.5	**56.7**	52.2	56.2	54.8	67.4	**62.2**	55.3	59.9
12	**Ours**	ResNet101	321 × 321	55.4	67.6	53.4	51.5	57.0	58.7	69.7	55.8	56.6	60.2
13	**Ours**	ResNet101	513 × 513	**55.7**	**68.5**	54.7	**53.2**	**58.0**	**60.8**	**70.6**	57.0	**57.5**	**61.5**

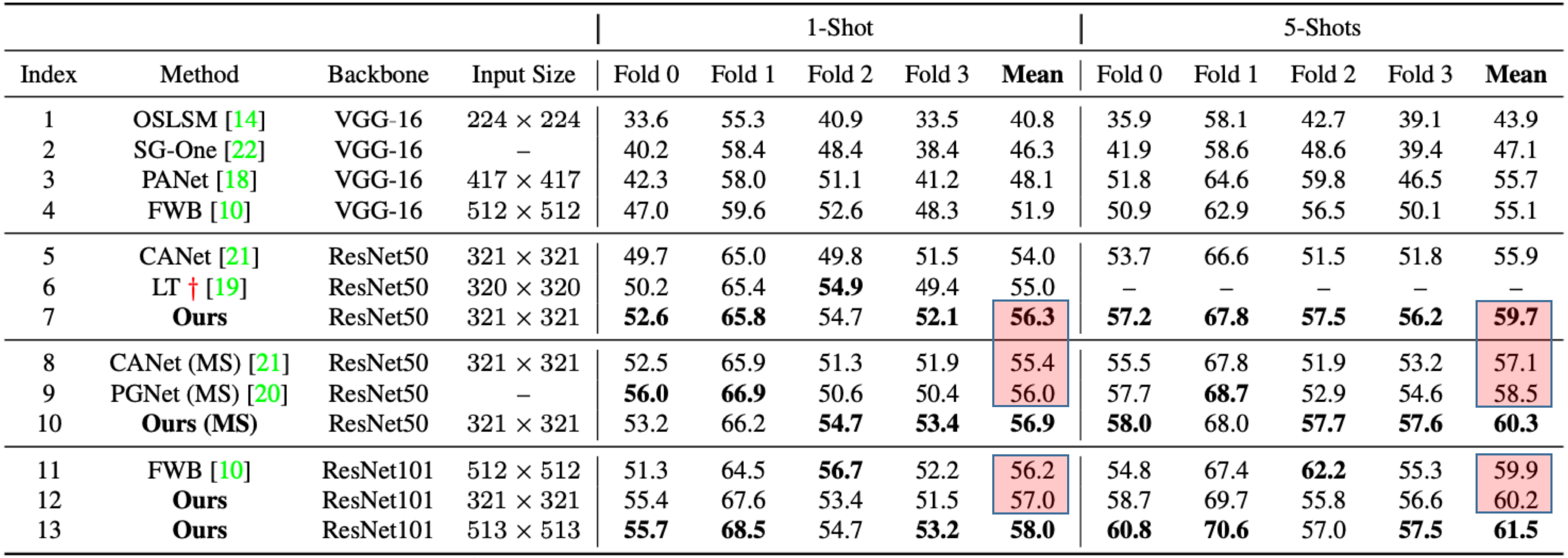
Experimental results on PASCAL-5i benchmark under Mean IoU metric.



			1-Shot					5-Shots				
Method	Backbone	Input Size	Fold 0	Fold 1	Fold 2	Fold 3	Mean	Fold 0	Fold 1	Fold 2	Fold 3	Mean
FWB [10]	ResNet101	512 × 512	17.0	18.0	21.0	**28.9**	21.2	19.1	21.5	24.0	30.1	23.7
Ours	ResNet101	513 × 513	**25.7**	**27.1**	**28.5**	25.6	**26.7**	**28.3**	**31.9**	**35.5**	**31.2**	**31.7**

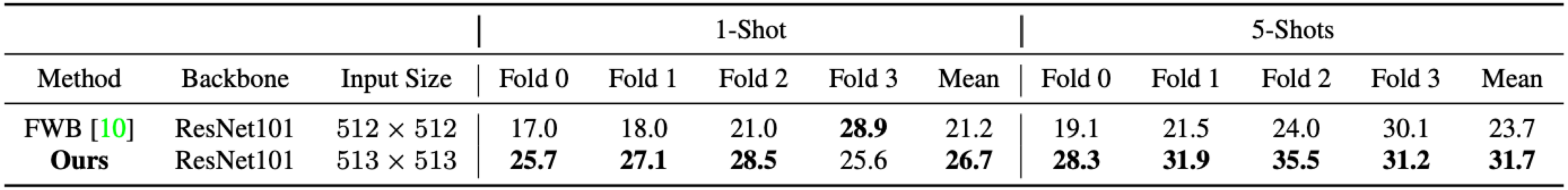
Experimental results on COCO-20i benchmark under Mean IoU metric.

Quantitative Results



				1-Shot					5-Shots				
Index	Method	Backbone	Input Size	Fold 0	Fold 1	Fold 2	Fold 3	**Mean**	Fold 0	Fold 1	Fold 2	Fold 3	**Mean**
1	OSLSM [14]	VGG-16	224 × 224	33.6	55.3	40.9	33.5	40.8	35.9	58.1	42.7	39.1	43.9
2	SG-One [22]	VGG-16	–	40.2	58.4	48.4	38.4	46.3	41.9	58.6	48.6	39.4	47.1
3	PANet [18]	VGG-16	417 × 417	42.3	58.0	51.1	41.2	48.1	51.8	64.6	59.8	46.5	55.7
4	FWB [10]	VGG-16	512 × 512	47.0	59.6	52.6	48.3	51.9	50.9	62.9	56.5	50.1	55.1
5	CANet [21]	ResNet50	321 × 321	49.7	65.0	49.8	51.5	54.0	53.7	66.6	51.5	51.8	55.9
6	LT † [19]	ResNet50	320 × 320	50.2	65.4	**54.9**	49.4	55.0	–	–	–	–	–
7	**Ours**	ResNet50	321 × 321	**52.6**	**65.8**	54.7	**52.1**	**56.3**	**57.2**	**67.8**	**57.5**	**56.2**	**59.7**
8	CANet (MS) [21]	ResNet50	321 × 321	52.5	65.9	51.3	51.9	55.4	55.5	67.8	51.9	53.2	57.1
9	PGNet (MS) [20]	ResNet50	–	**56.0**	**66.9**	50.6	50.4	56.0	57.7	**68.7**	52.9	54.6	58.5
10	**Ours (MS)**	ResNet50	321 × 321	53.2	66.2	**54.7**	**53.4**	**56.9**	**58.0**	68.0	**57.7**	**57.6**	**60.3**
11	FWB [10]	ResNet101	512 × 512	51.3	64.5	**56.7**	52.2	56.2	54.8	67.4	**62.2**	55.3	59.9
12	**Ours**	ResNet101	321 × 321	55.4	67.6	53.4	51.5	57.0	58.7	69.7	55.8	56.6	60.2
13	**Ours**	ResNet101	513 × 513	**55.7**	**68.5**	54.7	**53.2**	**58.0**	**60.8**	**70.6**	57.0	**57.5**	**61.5**

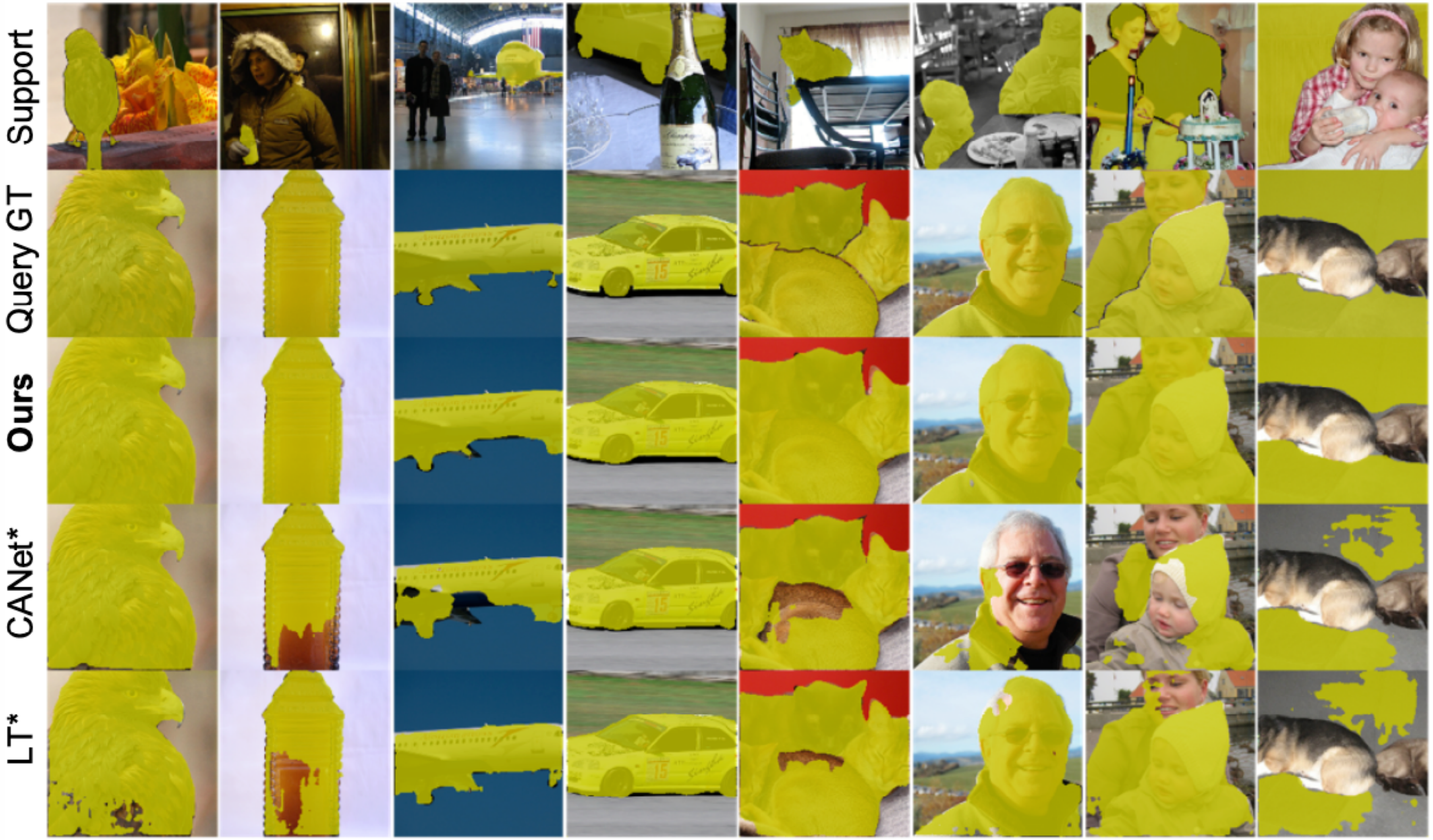
Experimental results on PASCAL-5i benchmark under Mean IoU metric.



			1-Shot					5-Shots				
Method	Backbone	Input Size	Fold 0	Fold 1	Fold 2	Fold 3	Mean	Fold 0	Fold 1	Fold 2	Fold 3	Mean
FWB [10]	ResNet101	512 × 512	17.0	18.0	21.0	**28.9**	21.2	19.1	21.5	24.0	30.1	23.7
Ours	ResNet101	513 × 513	**25.7**	**27.1**	**28.5**	25.6	**26.7**	**28.3**	**31.9**	**35.5**	**31.2**	**31.7**

Experimental results on COCO-20i benchmark under Mean IoU metric.

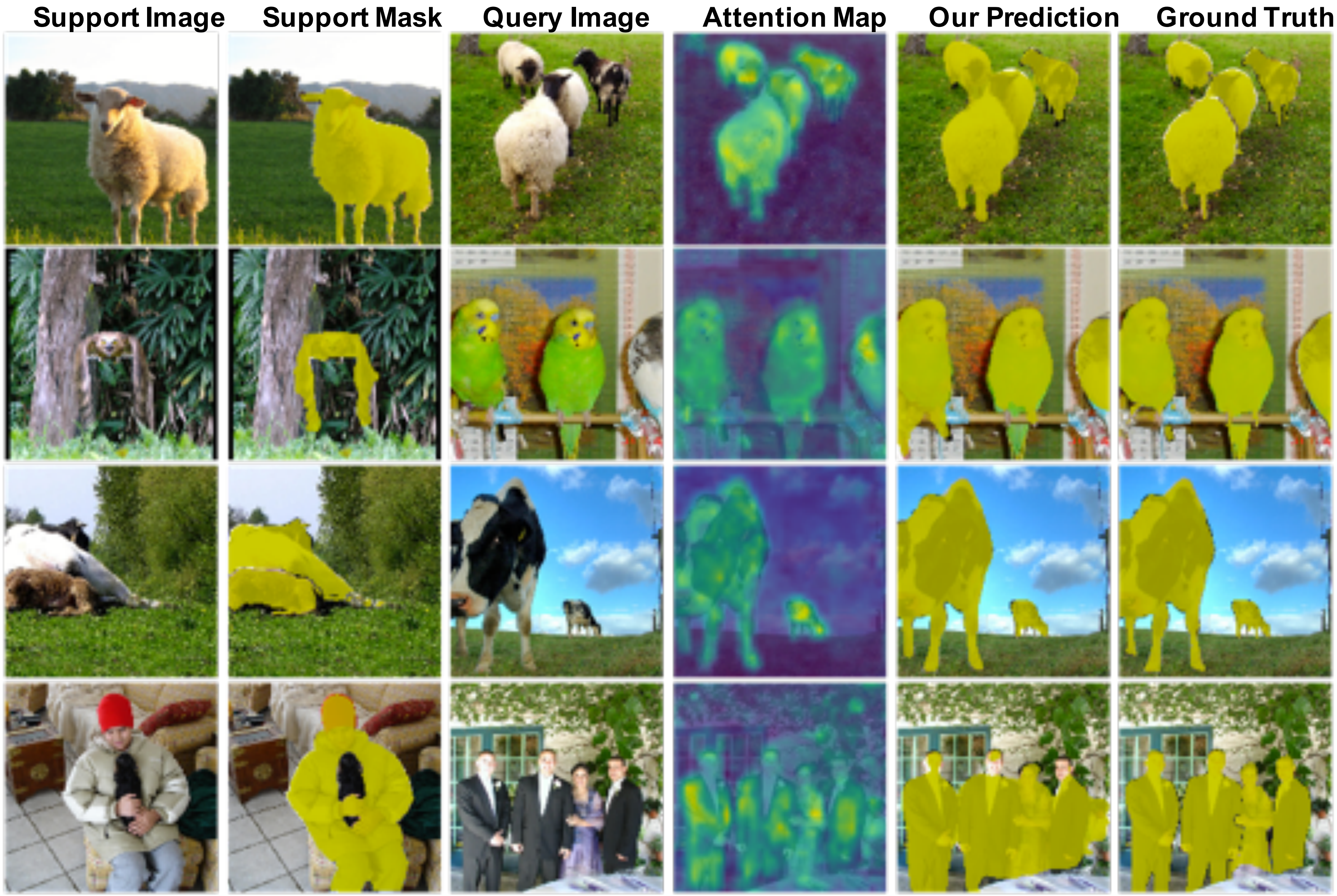
Qualitative Results



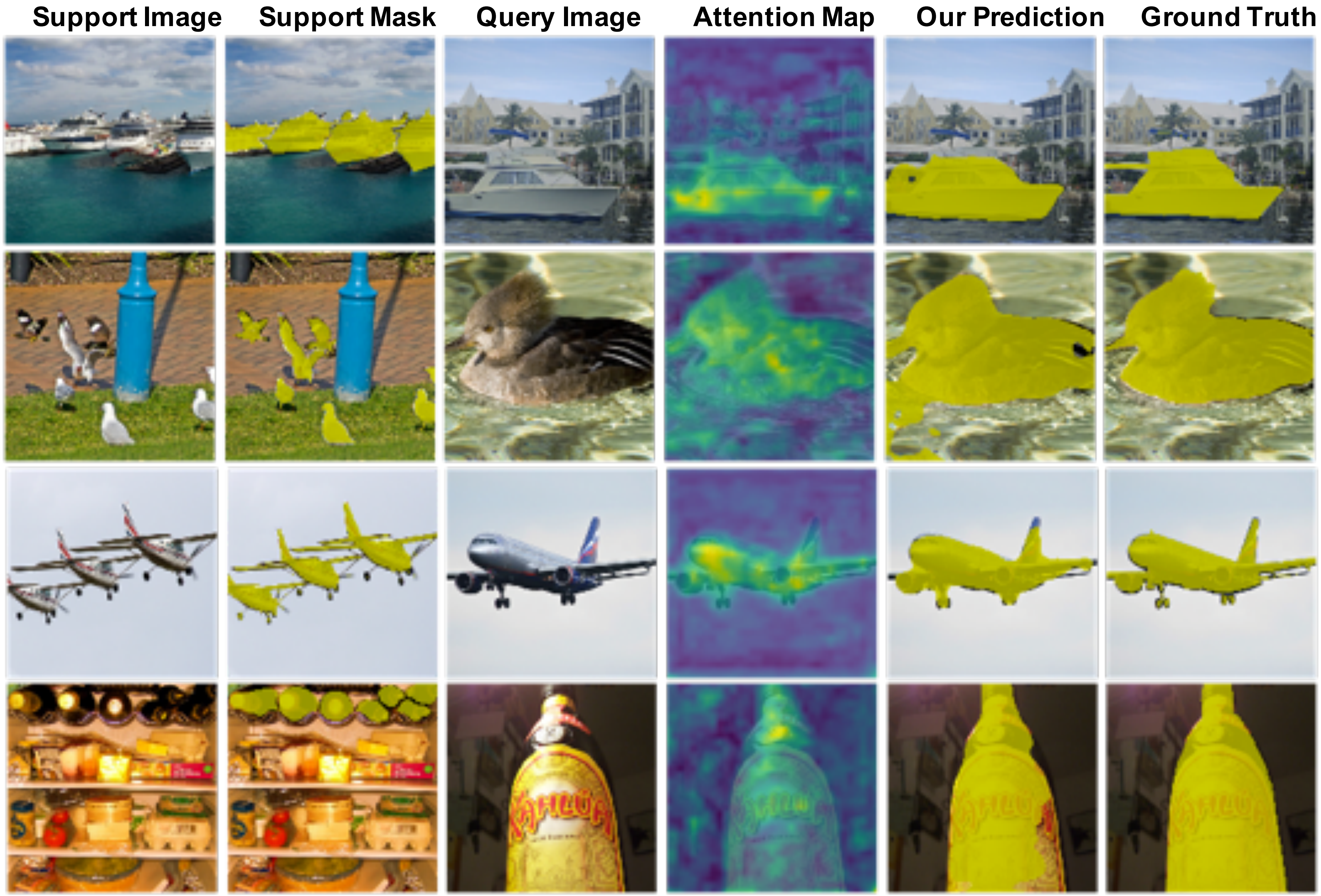
Example qualitative results selected from PASCAL 5i dataset

More Qualitative Results on Challenging Scenarios

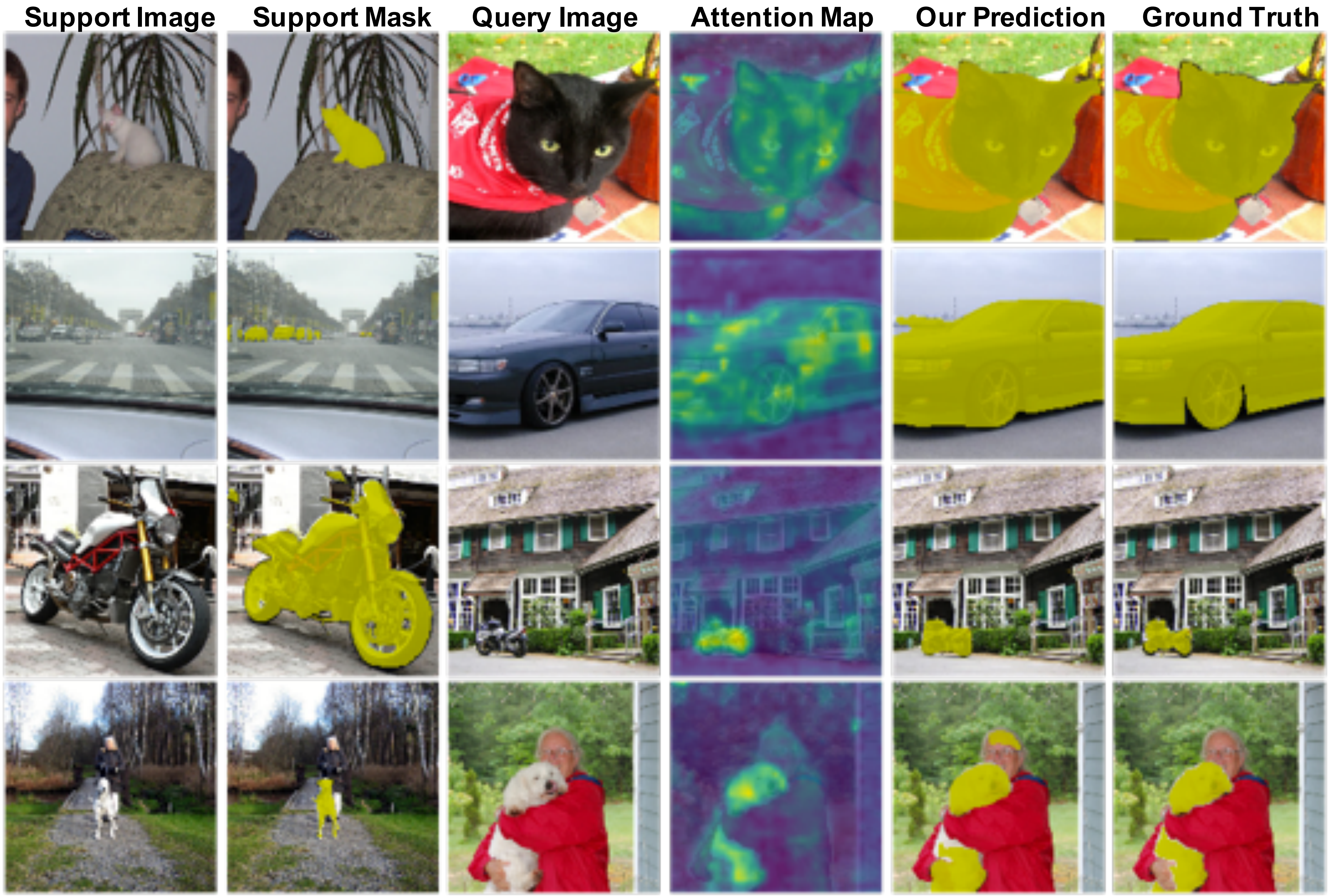
- **One-to-many Matching**: The support example has one object and the query image has multiple objects.
- **Many-to-one Matching**: The support example has multiple objects and the query image has only one object.
- **Small-to-large / Large-to-small Matching**: Objects in the support example are small while objects in the query image are large, or vice versa.
- **Change of Viewing Angles**: The viewing angle of an object in support image and query image has large variation.



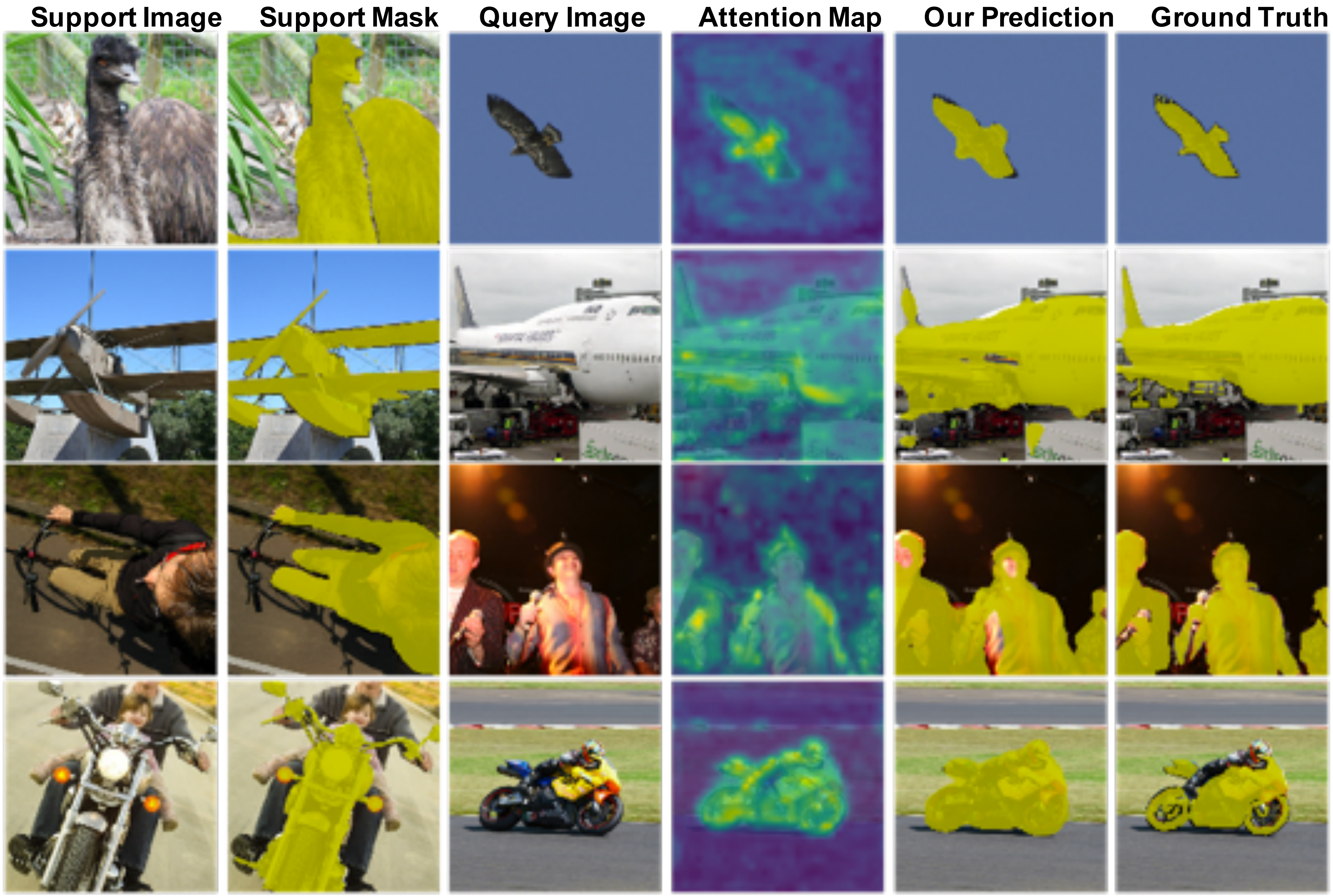
Example results under “one-to-many” matching scenarios.



Example results under “many-to-one” matching scenarios.

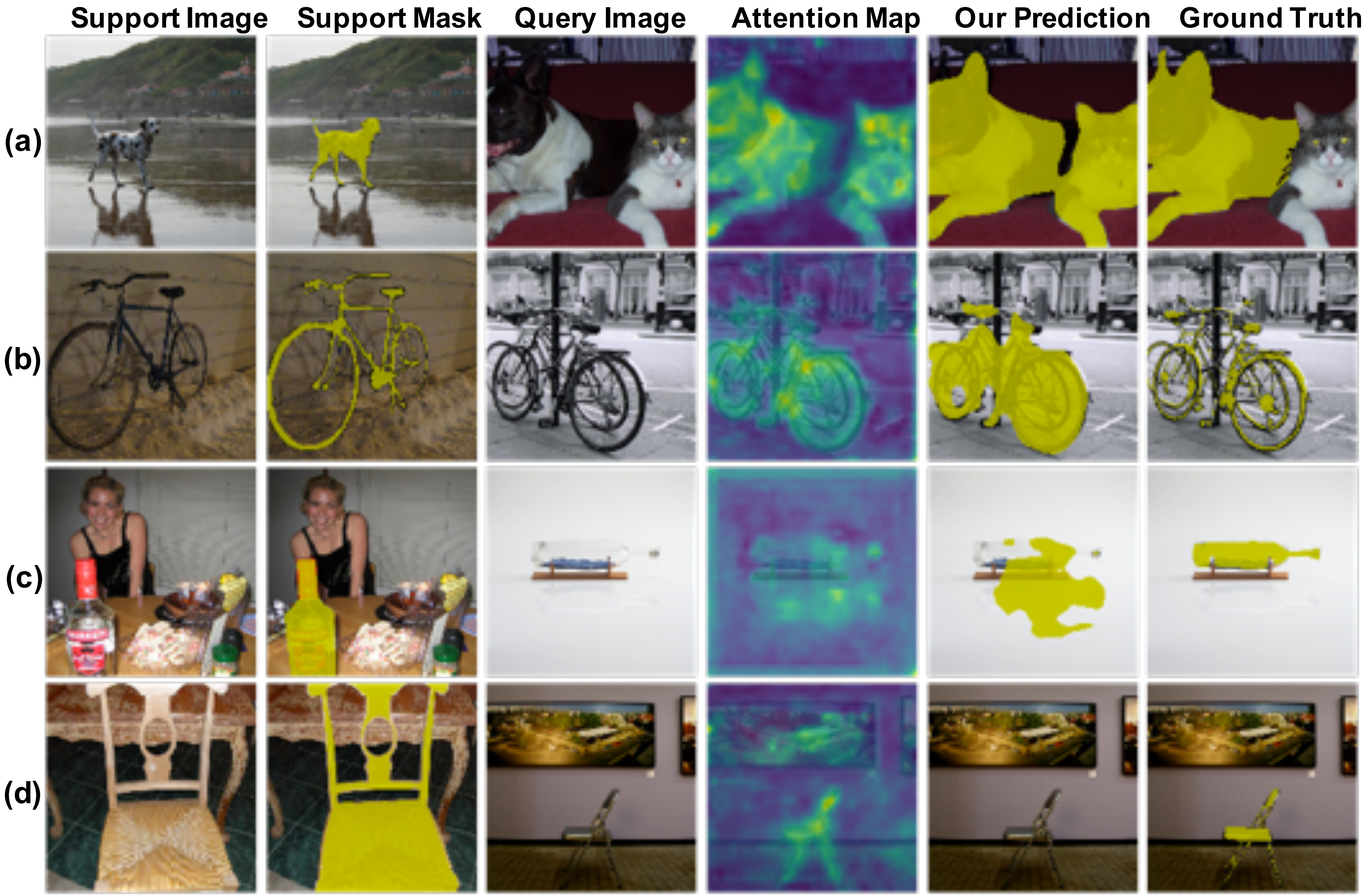


Example results when objects have large variations in object sizes.



Example results when objects have large variations in viewing angles.

Failure Cases



Thank You