Anti-aliasing Semantic Reconstruction for Few-Shot Semantic Segmentation

Binghao Liu[1] Yao Ding[1] Jianbin Jiao[1] Xiangyang Ji[2] Qixiang Ye[1*]
PriSDL, EECE, University of Chinese Academy of Sciences[1]
Department of Automation, Tsinghua University[2]
{liubinghao18, dingyao16}@mails.ucas.ac.cn xyji@tsinghua.edu.cn
{jiaojb, qxye}@ucas.ac.cn

Abstract

Encouraging progress in few-shot semantic segmentation has been made by leveraging features learned upon base classes with sufficient training data to represent novel classes with few-shot examples. However, this feature sharing mechanism inevitably causes semantic aliasing between novel classes when they have similar compositions of semantic concepts. In this paper, we reformulate few-shot segmentation as a semantic reconstruction problem, and convert base class features into a series of basis vectors which span a class-level semantic space for novel class reconstruction. By introducing contrastive loss, we maximize the orthogonality of basis vectors while minimizing semantic aliasing between classes. Within the reconstructed representation space, we further suppress interference from other classes by projecting query features to the support vector for precise semantic activation. Our proposed approach, referred to as anti-aliasing semantic reconstruction (ASR), provides a systematic yet interpretable solution for few-shot learning problems. Extensive experiments on PASCAL VOC and MS COCO datasets show that ASR achieves strong results compared with the prior works. Code will be released at github.com/Bibkiller/ASR.

1. Introduction

Over the past few years, we have witnessed the substantial progress of object detection and semantic segmentation [45, 46, 28, 48, 1, 14]. This can be attributed to convolutional neural networks (CNNs) with excellent representation capability and the availability of large datasets with concise mask annotations, especially. However, annotating a large number of object masks is expensive and infeasible in some scenarios (*e.g.*, computer-aided diagnosis systems). Few-shot semantic segmentation, which aims to generalize a model pre-trained on base classes of sufficient data to novel classes with only a few examples, has emerged as a promising technique.

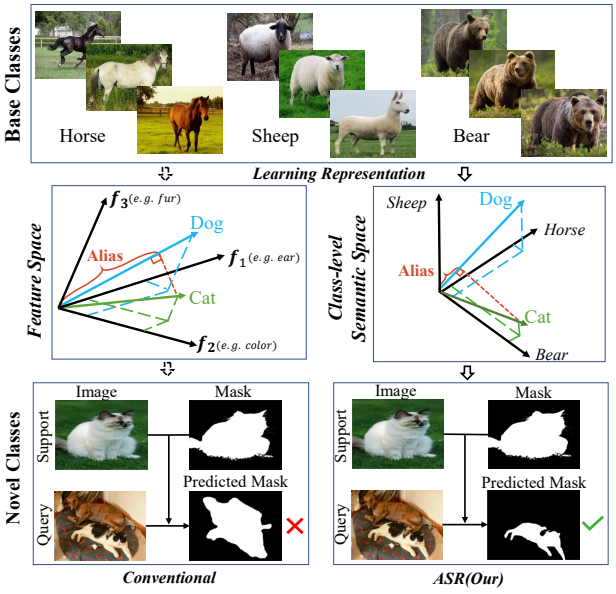


Figure 1. Comparison of conventional methods and our ASR method. While conventional methods represent novel classes (*e.g.*, cat and dog) within the feature space specified for base classes without considering the semantic aliasing, ASR implements semantic reconstruction by constructing a class-level semantic space where basis vectors are orthogonal and the semantic interference is reduced.

In few-shot segmentation, the generalization process is to utilize features learned upon base classes with sufficient training data to represent novel classes. However, for the overlapped semantics among features, the intricate many-to-many correspondence between features and classes inevitably causes semantic aliasing[1] between novel classes when they have similar compositions of semantic concepts. For example, a cat and a dog appear in the same query im-

[1] Semantic aliasing refers to an effect that causes classes to be indistinguishable due to the sharing of semantics among features.

*Corresponding Author.

age are confused because they correspond to the similar features of the base classes for bears and sheep, which results in false segmentation, Fig. 1(left).

In this paper, we reformulate the few-shot segmentation task as a semantic reconstruction problem and propose an anti-aliasing semantic reconstruction (ASR) approach. To fulfil semantic reconstruction, we first span a class-level semantic space. During the training phase, convolutional feature channels are categorized into channel groups, each of which is optimized for constructing a basis vector corresponding to a base class. This suppresses the semantic overlap between feature channels. We further introduce a contrastive loss to enhance the orthogonality of basis vectors and improve their representation capability. In the space, the semantic vectors of novel classes are represented by weighted basis vector reconstruction. Due to the potential class-level semantic similarity, the novel class will be reconstructed by its semantic-proximal base classes. In this way, novel classes inherit the orthogonality of base classes and are distinguishable, Fig. 1(middle right).

To suppress interfering semantics from the background or other classes within the same query image, we further propose the semantic filtering module, which projects query feature vectors to the reconstructed support vector. As the support images have precise semantics guided by the ground-truth annotations, the projection operation divorces interfering semantics, which facilities the activation of target object classes, Fig. 1(bottom right). In the metric learning framework, ASR implements semantic anti-aliasing between novel classes and within query images, providing a systematic solution for few-shot learning, Fig. 2. Such anti-aliasing can be analyzed from perspectives of vector orthogonality and sparse reconstruction, making ASR an interpretable approach.

The contributions of this study include:

- We propose a systematic and interpretable anti-aliasing semantic reconstruction (ASR) approach for few-shot semantic segmentation, by converting the base class features into a series of basis vectors for semantic reconstruction.
- We propose semantic span, which reduces the semantic aliasing between base classes for precise novel class reconstruction. Based on semantic span, we further propose semantic filtering, to eliminate interfering semantics within the query image.
- ASR improves the prior approaches with significant margins when applied to commonly used datasets. It also achieves good performance under the two-way few-shot segmentation settings.

2. Related Works

Semantic Segmentation. Benefiting from the superiority of fully convolutional networks, semantic segmentation [2, 39, 48] has progressed substantially in recent years. Relevant research has also provided some fundamental techniques, such as multi-scale feature aggregation [48] and atrous spatial pyramid pooling (ASPP) [2], which enhance few-shot semantic segmentation. However, these methods generally require large amounts of pixel-level annotations, which hinders their application in many real-world scenarios.

Few-shot Learning. While meta-learning [36, 27, 8, 15, 44, 38, 21] contributed important optimization methods and data augmentation [13, 35] aggregated performance, metric learning [32, 30, 11, 5] with prototype models [23, 4, 6, 41, 40, 18] represent the majority of few-shot learning approaches. In metric learning frameworks, prototypical models convert spatial semantic information of objects to convolutional channels. With prototypes, metric algorithms aim to obtain a high similarity score for similar sample pairs while a low similarity score for dissimilar pairs. For example, Ref. [3] replaced the fully connected layer with cosine similarity. Ref. [10] devises a few-shot visual learning system that performs well on both base and novel classes. DeepEMD [41] proposed the structural distance between dense image representations. Extra margin constraints [20, 17] are absorbed into metric learning to further adjust the inter-class diversity and intra-class variance. Despite the popularity of metric learning, the semantic aliasing issue caused by the feature sharing mechanism is unfortunately ignored.

Few-shot Segmentation. Early methods generally utilized a parametric module, which uses features learned through support image(s) to segment the query image. In [26] support features were concatenated with the query image to activate features within object regions for segmentation. PGNet [42] and DAN [33] tackled semantic segmentation with graphs and used graph reasoning to propagate label information to the query image.

Following few-shot classification, prototype vectors have been used as semantic representation across feature channels. In [47], masked average pooling was utilized to squeeze foreground information within the support image(s) to prototype vectors. CANet [43] consisted of a two-branch model which performs feature comparison between the support image(s) and the query image guided by prototypes. PANet [34] offered highly representative prototypes for each semantic class and performs segmentation over the query image based on pixel-wise matching. CRNet [22] proposed a cross-reference mechanism to concurrently make predictions for both the support image(s) and the query image, enforcing co-occurrence of objects and thereby improving the semantic transfer.

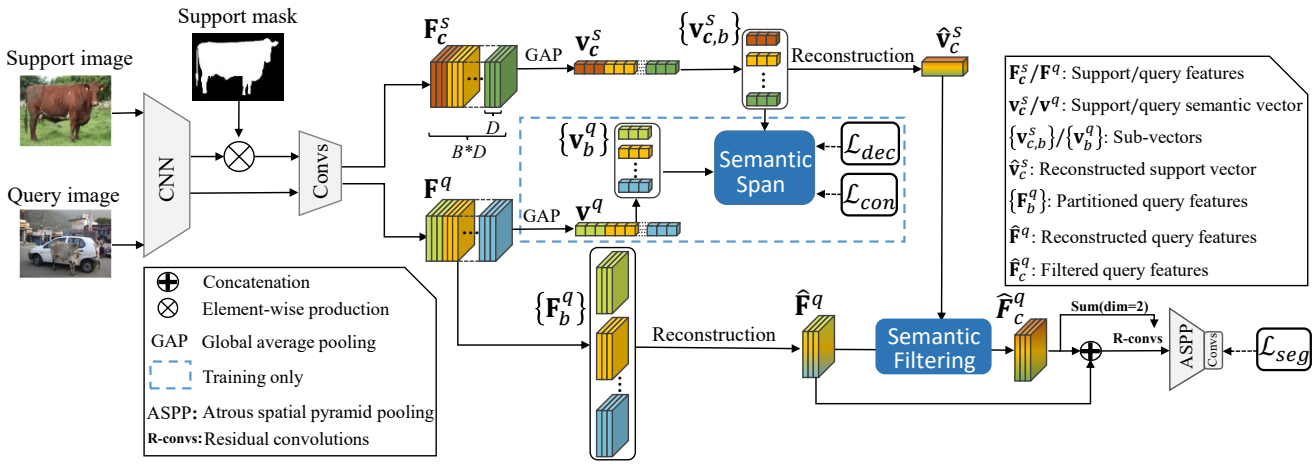


Figure 2. Few-shot segmentation flowchart with Anti-aliasing Semantic Reconstruction (ASR). The flowchart defines a metric learning framework consisting of a support branch (upper) and a query branch (lower), where reconstruction, semantic span, and semantic filtering modules are plugged. While semantic span reduces the semantic aliasing between base classes driven by contrastive loss, semantic filtering aims to suppress interfering semantics within the query image. (Best viewed in color)

PMMs [37] and PPNet [24] proposed to decompose objects into parts and represent such parts with mixed prototype vectors to counter semantic mixing. Despite the aforementioned progress, existing methods remain ignorant of the semantic aliasing issue, which causes false (or missing) segmentation of object parts. SST [49] and SimProp [9] respectively introduced self-supervised finetuning and similarity propagation, which leverage the category-specific semantic constraints to reduce semantic aliasing. However, without considering the orthogonality of base class features, they remain challenged by the semantic aliasing issue.

3. The Proposed Method

3.1. Problem Definition

Few-shot semantic segmentation aims to learn a model ($e.g.$, a network) which can generalize to previously unseen classes. Given two image sets $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$, classes in $\mathcal{D}_{novel}$ do not appear in $\mathcal{D}_{base}$, it requires to train the feature representation on $\mathcal{D}_{base}$ (which has sufficient data) and test on $\mathcal{D}_{novel}$ (which has only a few annotations). Both $\mathcal{D}_{base}$ and $\mathcal{D}_{novel}$ contain several episodes, each of which consists of a support set $(\mathbf{A}_i^s, \mathbf{M}_i^s)_{i=1}^K$ and a query set $(\mathbf{A}^q, \mathbf{M}^q)$, where $K, \mathbf{A}_i^s, \mathbf{M}_i^s, \mathbf{A}^q$ and $\mathbf{M}^q$ respectively represent the shot number, the support image, the support mask, the query image, and the query mask. For each training episode, the model is optimized to segment $\mathbf{A}^q$ driven by the segmentation loss $\mathcal{L}_{seg}$. Segmentation performance is evaluated on $\mathcal{D}_{novel}$ across all the test episodes.

3.2. Semantic Reconstruction Framework

We propose a semantic reconstruction framework, where the semantics of novel classes are explicitly reconstructed by those of base classes, Fig. 2. Given support and query images, after extracting convolutional features through a CNN, the ground-truth mask is multiplied with support features in a pixel-wised fashion to filter out background features [47, 43, 22, 37, 24]. With a convolutional block we reduce the number of feature channels and obtain support features $\mathbf{F}_c^s \in \mathbb{R}^{H\times W\times(B\times D)}$ and query features $\mathbf{F}^q \in \mathbb{R}^{H\times W\times(B\times D)}$, where $H \times W$, B, and D respectively denote the size of feature maps, base class number, and feature channel number. During training phase, c denotes the base class. And c denotes the novel class during testing phase. The convolutional block consists of pyramid convolution layers, which captures features from coarse to fine. To explicitly encode class-related semantics, we averagely partition the feature channels to B groups, corresponding to B base classes. The grouped features $\mathbf{F}_c^s$ and $\mathbf{F}^q$ are further spatially squeezed into two vectors $\mathbf{v}_c^s$ and $\mathbf{v}^q$, termed semantic vectors, by global average pooling, Fig. 2.

Corresponding to B base classes, the semantic vectors $\mathbf{v}_c^s$ and $\mathbf{v}^q$ consists of B sub-vectors $\{\mathbf{v}_{c,b}^s\}_{\{b=1,2,\dots,B\}} \in \mathbb{R}^D$ and $\{\mathbf{v}_b^q\}_{\{b=1,2,\dots,B\}} \in \mathbb{R}^D$. During the training phase, the sub-vectors are used to construct basis vectors in the B-dimensional class-level semantic space by the semantic span module, as explained in Section 3.3, Fig. 2. In the space, a basis vector ($\mathbf{v}_b$) corresponding to the b-th base class is defined as $\mathbf{v}_b = \mathbf{v}_{c,b}^s/||\mathbf{v}_{c,b}^s|| = \mathbf{v}_b^q/||\mathbf{v}_b^q||$. In the inference phase, the semantic vector for the c-th class in support branch can be linearly reconstructed [16], as

$$\hat{\mathbf{v}}_c^s = \sum_{b=1}^{B} w_{c,b}^s \cdot \mathbf{v}_b, \tag{1}$$

where $\hat{\mathbf{v}}_c^s$ denotes the reconstructed support semantic vector (reconstructed support vector for short), and

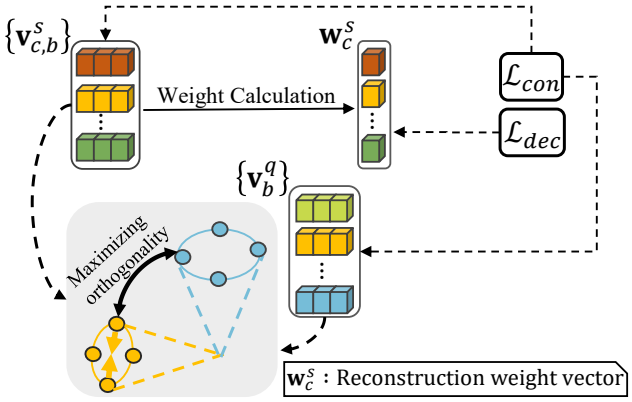


Figure 3. Implementation and illustration of semantic span, which constructs basis vectors and enhances the orthogonality of them. Semantic span is driven by the semantic decoupling loss and constrastive loss. (Best viewed in color)

$w^s_{c,b}$ is the b-th element of the weight vector $\mathbf{w}^s_c = softmax([||\mathbf{v}^s_{c,1}||, ||\mathbf{v}^s_{c,2}||, \cdots, ||\mathbf{v}^s_{c,B}||])$. Large $w^s_{c,b}$ indicates that the basis vector, whose corresponding base class have strong similarity with the c-th class, contributes much to the reconstruction.

Consistently, given a query image, the corresponding query features $\mathbf{F}^q$ are reconstructed by regarding each location on the feature maps as a feature vector. Each location of feature map is reconstructed as $\hat{\mathbf{F}}^q(x,y) = \sum_{b=1}^{B} \mathbf{W}^q_b(x,y) \cdot \mathbf{v}_b$, where (x,y) denotes the coordinates of pixels on the feature map, and $\mathbf{W}^q_b(x,y)$ is defined as the norm of sub-vector $\mathbf{F}^q_b(x,y)$. Considering that the query image contains objects not only belonging to the target class but also other classes, we exploit the semantic filtering module, as illustrated in Section 3.4, to filter out the interfering components in the reconstructed query features for the c-th target class semantic segmentation.

3.3. Semantic Span

Within the origin feature space, when base class features are close to each other, there could be semantic aliasing among novel classes. To minimize semantic aliasing, we propose to span a class-level semantic space in the training phase. To construct a group of basis vectors which tends to be orthogonal and representative, we propose the semantic span module (semantic span for short). As shown in Fig. 3, the semantic span is driven by two loss functions, $i.e.$, semantic decoupling and contrastive losses.

On the one hand, the semantic span targets at constructing basis vectors by regularizing the feature maps so that each group of features is correlated to a special object class. To fulfill this purpose, we propose the following semantic decoupling loss, as

$$\mathcal{L}_{dec} = \log(1 + e^{-\mathbf{w}^s_c \cdot \mathbf{y}}), \tag{2}$$

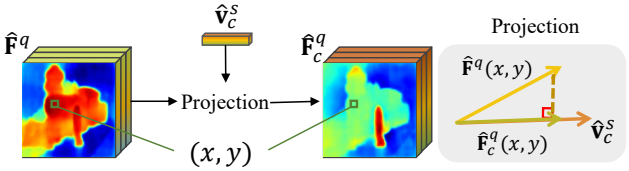


Figure 4. Semantic filtering with vector projection to suppress interfering semantics within the image. (Best viewed in color)

where $\mathbf{w}^s_c$ denotes the reconstruction weight vector, and $\mathbf{y} \in \mathbb{R}^B$ denotes the one-hot class label of a support image. Obviously, minimizing $\mathcal{L}_{dec}$ is equivalent to maximize the reconstruction weights related to the specific class ($e.g.$, the c-th class), while minimizing those unrelated to it. This defines a soft manner converting a group of features correlated to the semantics of the specific class ($e.g.$, the c-th class) to its corresponding basis vector in the class-level semantic space.

On the other hand, the semantic span targets at further enhancing orthogonality of basis vectors, which improve the quality of novel class reconstruction. In details, subvectors in $\{\mathbf{v}^s_{c,b}\}_{\{b=1\ldots B\}} \cup \{\mathbf{v}^q_b\}_{\{b=1\ldots B\}}$ belonging to different classes are expected to be orthogonal to each other while those corresponding to the same base classes, $e.g.$, $\mathbf{v}^s_{c,b}$ and $\mathbf{v}^q_b$, are expected to have a small vector angle. These two objectives are simultaneously achieved by minimizing the contrastive loss defined as

$$\mathcal{L}_{con} = \frac{e^{1+\sum_{b\neq b'} |cos<\mathbf{v}^s_{c,b}, \mathbf{v}^q_{b'}>|}}{e^{|cos<\mathbf{v}^s_{c,b}, \mathbf{v}^q_b>|}}, \tag{3}$$

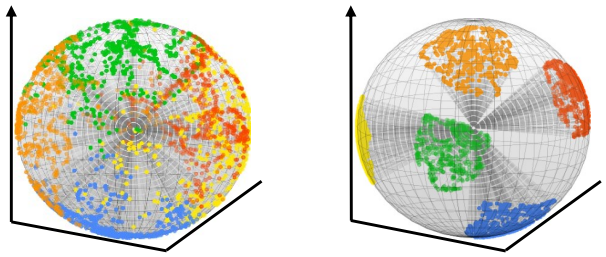
where $cos < \cdot >$ denotes the $Cosine$ distance metric of two vectors. In summary, the final loss of the semantic reconstruction framework is defined as:

$$\mathcal{L} = \alpha\mathcal{L}_{dec} + \beta\mathcal{L}_{seg} + \gamma\mathcal{L}_{con}, \tag{4}$$

where α, β and γ are weights of the loss functions. Note that $\mathcal{L}_{con}$ is calculated during the later stage of training phase.

3.4. Semantic Filtering

When multiple objects from different classes exist in the same query image, the reconstructed features of the query image contains components of all these classes. To pick out objects belonging to the target class and suppress interfering semantics, $i.e.$, divorcing the semantics related to the background or objects from other classes, we propose a semantic filtering module. Moreover, owing to that the reconstructed vectors of different classes are non-collinear, the semantic filtering module is implemented by projecting query feature vectors to the reconstructed support vector, as shown in Fig. 4. This is also based on the fact that the reconstructed support vector has precise semantics because



(w/o) semantic span (w) semantic span

Figure 5. While the semantic vectors without semantic span tends to be mixed up in the semantic space, those with semantic span repel one another towards orthogonality. (Best viewed in color)

the corresponding features have been multiplied with the ground-truth mask, as shown in Fig. 2.

On the support branch, we follow Eq. 1 to reconstruct the support features using basis vectors and obtain the reconstructed support vector $\hat{\mathbf{v}}_c^s$. On the query branch, we reconstruct each feature vector $\mathbf{F}^q(x,y)$ on the feature maps in the same way and obtain the reconstructed features $\hat{F}^q$. We then project $\hat{\mathbf{F}}^q$ to $\hat{\mathbf{v}}_c^s$ to calculate filtered features as

$$\hat{\mathbf{F}}_c^q(x,y) = \frac{\hat{\mathbf{F}}^q(x,y)\cdot\hat{\mathbf{v}}_c^s}{||\hat{\mathbf{v}}_c^s||}\cdot\frac{\hat{\mathbf{v}}_c^s}{||\hat{\mathbf{v}}_c^s||}, \tag{5}$$

where (x,y) denotes the coordinates of pixels on the feature maps. The intuitive effects of the filter operation are displayed in Fig. 4, which illustrates that the support branch guides the query branch more effectively. The filtered query features $\hat{\mathbf{F}}_c^q$ are further enhanced by a residual convolutional module with iterative refinement optimization and fed to Atrous Spatial Pyramid Pooling (ASPP) to predict the segmentation mask, Fig. 2. For the residual convolutional module, we replace the history mask in CANet [43] with the squeezed $\hat{\mathbf{F}}_c^q$.

3.5. Interpretable Analysis

ASR can be analyzed from the perspectives of vector orthogonality and sparse reconstruction. Without loss of generality, we take the two-dimensional space as an example. Denote $\mathbf{v_1}, \mathbf{v}_2 \in \mathbb{R}^2$ two unit basis vectors ($\|\mathbf{v}_1\| = 1, \|\mathbf{v}_2\| = 1$), which span the space. Denote θ as the angle between $\mathbf{v}_1$ and $\mathbf{v}_2$, and $cos\,\theta = \mathbf{v}_1\mathbf{v}_2$. According to the properties of linear algebra [16], any vectors, $e.g.$, $\mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^2$, in the spanned space can be linearly reconstructed as

$$\begin{cases}\mathbf{u}_1 = C_1(w_{11}\mathbf{v}_1 + w_{12}\mathbf{v}_2), \forall\mathbf{u}_1 \in \mathbb{R}^2\\ \mathbf{u}_2 = C_2(w_{21}\mathbf{v}_1 + w_{22}\mathbf{v}_2), \forall\mathbf{u}_2 \in \mathbb{R}^2\end{cases} \tag{6}$$

where $w_{11}, w_{12}, w_{21}, w_{22} \in [0,1]$ are reconstruction weights which feed the linear constraints: $w_{11}+w_{12}=1.0$ and $w_{21}+w_{22}=1.0$. C_1 and C_2 are scaling constants. The

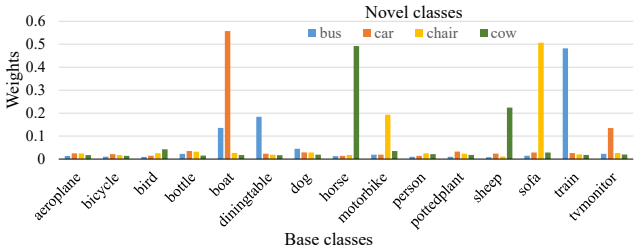


Figure 6. Sparse reconstruction weights of novel classes using the base classes. (Best viewed in color)

cosine similarity between $\mathbf{u}_1$ and $\mathbf{u}_2$ is computed as

$$\begin{aligned} cos <\mathbf{u}_1,\mathbf{u}_2> &= \frac{\mathbf{u}_1\mathbf{u}_2}{|\mathbf{u}_1|\,|\mathbf{u}_2|}\\ &= (w_{11}\mathbf{v}_1 + w_{12}\mathbf{v}_2)(w_{21}\mathbf{v}_1 + w_{22}\mathbf{v}_2)\\ &= w_{11}w_{21} + (w_{11}w_{22} + w_{21}w_{12})\mathbf{v}_1\mathbf{v}_2 + w_{12}w_{22}\\ &= w_{11}w_{21} + (1-w_{11})(1-w_{21})\\ &\quad + [w_{11}(1-w_{21}) + w_{21}(1-w_{11})]cos\,\theta\\ &= 1 + (w_{11} + w_{21} - 2w_{11}w_{21})(cos\,\theta - 1), \end{aligned} \tag{7}$$

where $(w_{11} + w_{21} - 2w_{11}w_{21}) \in [0,1]$ and $(cos\,\theta - 1) \in [-1,0]$.

Orthogonality. To reduce semantic aliasing of any two novel classes, the angle between their semantic vectors, $\mathbf{u}_1$ and $\mathbf{u}_2$, should be large, known as to reduce $cos < \mathbf{u}_1,\mathbf{u}_2 >$. According to the last line of Eq. 7, to obtain a small $cos < \mathbf{u}_1,\mathbf{u}_2 >$, the term $cos\,\theta$ should approach to 0, which means that the angle between the basis vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ is large, which implies the orthogonality of basis vectors. The proposed ASR approach satisfies the orthogonality by introducing the semantic span module. As shown in Fig.5, the statistical visualization results over base classes validate the orthogonality.

Sparse Reconstruction. Refer to the last line of Eq. 7, the another manner to reduce the $cos < \mathbf{u}_1,\mathbf{u}_2 >$ is enlarging the term $(w_{11} + w_{21} - 2w_{11}w_{21})$. According to the function characteristics, $(w_{11} + w_{21} - 2w_{11}w_{21})$ reaches its maximum when $|w_{11}\text{-}w_{21}|$ approaches to 1.0, which contains underlying conditions that $|w_{11}\text{-}w_{12}|$ and $|w_{21}\text{-}w_{22}|$ approach to 1.0 due to the linear constraints. This illustrates that to further aggregate the capability of anti-aliasing and guarantee the discrimination of novel classes, the reconstruction weights for novel classes should be differential and sparse. ASR satisfies these requirements due to the potential class-level semantic similarity according to the statistical results shown in Fig. 6. Meanwhile, nonzero weights over multiple basis classes other than the dominate one enable ASR to distinguish novel classes from base classes.

Backbone	Method	1-shot					5-shot				
		Pascal-5^0	Pascal-5^1	Pascal-5^2	Pascal-5^3	Mean	Pascal-5^0	Pascal-5^1	Pascal-5^2	Pascal-5^3	Mean
VGG16	OSLSM [29]	33.60	55.30	40.90	33.50	40.80	35.90	58.10	42.70	39.10	43.95
	co-FCN [26]	36.70	50.60	44.90	32.40	41.10	-	-	-	-	-
	SG-One [47]	40.20	58.40	48.40	38.40	46.30	41.90	58.60	48.60	39.40	47.10
	PANet [34]	42.30	58.00	51.10	41.20	48.10	51.80	64.60	59.80	46.05	55.70
	FWB [25]	47.04	59.64	52.61	48.27	51.90	50.87	62.86	56.48	50.09	55.08
	PFENet [31]	56.90	68.20	54.40	52.40	58.00	59.00	69.10	54.80	52.90	59.00
	RPMMs [37]	47.14	65.82	50.57	48.54	53.02	50.00	66.46	51.94	47.64	54.01
	SST [49]	50.90	63.00	53.60	49.60	54.30	52.50	64.80	59.50	51.30	57.00
	ASR (ours)	49.19	65.41	52.58	51.32	54.63	52.52	66.51	54.98	53.85	56.97
	ASR* (ours)	50.21	66.35	54.26	51.81	55.66	53.68	68.49	55.03	54.78	57.99
Resnet50	CANet [43]	52.50	65.90	51.30	51.90	55.40	55.50	67.80	51.90	53.20	57.10
	PGNet [42]	56.00	66.90	50.60	50.40	56.00	57.70	68.70	52.90	54.60	58.50
	CRNet [22]	-	-	-	-	55.70	-	-	-	-	58.80
	PPNet [24]	48.58	60.58	**55.71**	46.47	52.84	58.85	68.28	**66.77**	**57.98**	**62.97**
	SimPropNet [9]	54.86	67.33	54.52	52.02	57.19	57.20	68.50	58.40	56.05	60.04
	DAN [33]	-	-	-	-	57.10	-	-	-	-	59.50
	PFENet [31]	**61.70**	69.50	55.40	**56.30**	**60.80**	**63.10**	70.70	55.80	57.90	61.90
	RPMMs [37]	55.15	66.91	52.61	50.68	56.34	56.28	67.34	54.52	51.00	57.30
	ASR (ours)	53.81	**69.56**	51.63	52.76	56.94	56.17	70.56	53.89	53.38	58.50
	ASR* (ours)	55.23	**70.36**	53.38	53.66	58.16	59.38	**71.85**	56.87	55.72	60.96

Table 1. Mean-IoU performance of 1-way 1-shot and 5-shot segmentation on Pascal-5^i. ASR* denotes ASR with multi-scale evaluation.

Method	1-shot					5-shot				
	COCO-20^0	COCO-20^1	COCO-20^2	COCO-20^3	Mean	COCO-20^0	COCO-20^1	COCO-20^2	COCO-20^3	Mean
FWB [25]	16.98	17.98	20.96	28.85	21.19	19.13	21.46	23.93	30.08	23.65
PFENet [31]	**34.30**	33.00	32.30	30.10	32.40	**38.50**	38.60	**38.20**	34.30	**37.40**
SST [49]	-	-	-	-	22.20	-	-	-	-	31.30
DAN [33]	-	-	-	-	24.40	-	-	-	-	29.60
RPMMs [37]	29.53	**36.82**	28.94	27.02	30.58	33.82	**41.96**	32.99	33.33	35.52
ASR (ours)	29.89	34.98	31.86	**33.51**	**32.56**	31.26	37.86	33.47	**35.21**	34.35
ASR* (ours)	30.62	36.73	**32.68**	**35.35**	**33.85**	33.12	39.51	34.16	**36.21**	**35.75**

Table 2. Mean-IoU performance of 1-shot and 5-shot semantic segmentation on COCO-20^i. FWB and PFENet use the ResNet101 backbone while other approaches use the ResNet50 backbone. ASR* denotes ASR with multi-scale evaluation.

4. Experiments

In this section, we first describe the experimental settings. We then report the performance of ASR and compare it with state-of-the-art methods. We finally present ablation studies with experimental analysis and test the effectiveness of ASR on other few-shot learning tasks.

4.1. Experimental Settings

Datasets. The experiments are conducted on PASCAL VOC 2012 [7] and MS COCO [19] datasets. We combine the PASCAL VOC 2012 with SBD [12] and separate the combined dataset into four splits. The cross-validation method is used to evaluate the proposed approach by sampling one split as test categories $\mathcal{C}_{test} = 4i + 1, \ldots, 4i + 5$, where i is the index of a split. The remaining three splits are set as base classes for training. The reorganized dataset is termed as Pascal-5^i [33, 37]. Following the settings in [25, 33, 37] we construct the COCO-20^i dataset. MS COCO is divided into four splits, each of which contains 20 categories. We follow the same scheme for training and evaluation as on the Pascal-5^i. The category labels for the four splits are included in the supplementary material. For each split, 1000 pairs of support and query images are randomly selected for performance evaluation.

Method	1-shot	5-shot
SG-One [47]	63.9	65.9
PANet [34]	66.5	70.7
CANet [43]	66.2	69.6
PGNet [42]	69.9	70.5
CRNet [22]	66.8	71.5
PFENet [31]	**73.30**	73.90
DAN [33]	71.90	72.30
PPNet [24]	69.19	**75.76**
ASR (ours)	71.33	72.51
ASR* (ours)	72.86	74.12

Table 3. Comparison of FB-IoU performance on Pascal-5^i. ASR* denotes ASR with multi-scale evaluation.

Training and Evaluation. We use CANet [43] without attention modules as the baseline. In training, we set the learning rate as 0.00045. The segmentation model (network) is trained for 200000 steps with the poly de-

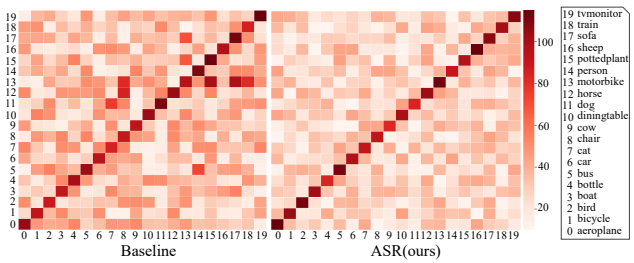


Figure 7. Confusion matrices of baseline method and the proposed ASR approach. (Best viewed in color)

scent training strategy and the stochastic gradient descent (SGD) optimizer. Several data augmentation strategies including normalization, horizontal flipping, gaussian filtering, random cropping, random rotation and random resizing are used. We adopt both the single-scale and multi-scale [43, 42, 22] evaluation strategies during testing. Our approach is implemented upon the PyTorch 1.3 and run on Nvidia Tesla V100 GPUs.

Evaluation Metric. Following [34, 25, 43], we use the mean Intersection over Union (mIoU) and binary Intersection over Union (FB-IoU) as the performance evaluation metrics. The mIoU calculates the per-class foreground IoU and averages the IoU for all classes to obtain the final evaluation metric. The FB-IoU calculates the mean of foreground IoU and background IoU over all images regardless of category. For category k, IoU is defined as $IoU_k = TP_k/(TP_k + FP_k + FN_k)$, where the TP_k, FP_k and FN_k are the number of true positives, false positives and false negatives in segmentation masks. mIoU is the average of IoUs for all the test categories and FB-IoU is the average of IoUs for all the test categories and the background. We report the segmentation performance by averaging the mIoUs on the four cross-validation splits.

4.2. Segmentation Performance

PASCAL VOC. In Table 1, we report the performance on Pascal VOC. ASR outperforms the prior methods with significant margins. Under 1-shot settings, with a VGG16 backbone, it respectively outperforms RPMMs [37] and SST [49] by 2.64% and 1.36%. Under the 1-shot settings, with a ResNet50 backbone, ASR outperforms CANet [43] and RPMMs [37] method by 2.76% and 1.82%. Under the 5-shot settings, ASR is comparable to the state-of-the-art method. It is worth mentioning that the SST and PPNet used additional k-shot fusion strategies while ASR uses a simple averaging strategy to get five-shot results. In Table 3, ASR is compared with state-of-the-art approaches with respect to FB-IoU. FB-IoU calculates the mean of foreground IoU and background IoU over images regardless of the categories, which reflects how well the full object extent is activated. ASR is on par with the compared methods, if not outperforms.

Semantic Reconst.	Semantic Span	Semantic Filter.	mIoU
			54.95
✓			53.26
		✓	53.12
✓	✓		55.98
✓	✓	✓	**58.64**

Table 4. Ablation of ASR modules. The baseline is CANet.

Concat.	Cosine	Conv.	Projection	mIoU
✓				58.21
	✓			57.78
		✓		58.32
			✓	**58.64**

Table 5. Comparison of semantic filtering strategies. Concat., Cosine, Conv., Projection denote vector concatenation, cosine similarity, convolutional operation, and vector projection, respectively.

MS COCO. In Table 2, we report the segmentation performance on MS COCO. ASR outperforms the prior methods in most settings. Particularly under the 1-shot setting, it improves RPMMs [37] by 3.27%. Under 5-shot setting, it improves DAN [33] by 6.15%, which are significant margins. For the MS COCO dataset with larger semantic aliasing for the more object categories, semantic reconstruction demonstrated larger advantages. For the larger object category number, we construct a space using more orthogonal basis vectors, which have stronger ability of representation and discrimination. According to Section 3.5, semantic aliasing among novel classes is suppressed effectively. That is why ASR achieves larger performance gains on the MS COCO dataset.

4.3. Visualization Analysis

We sampled 4000 images from 20 classes in PASCAL VOC, and drew the confusion matrix according to the segmentation results, Fig. 7. ASR effectively reduce semantic aliasing among classes. We further visualize segmentation results and compare them with baseline, Fig. 8. Based on the anti-aliasing representation of novel classes and semantic filtering, ASR reduces the false positive segmentation caused by interfering semantics within the query images.

4.4. Ablation Studies

Semantic Span. In Table 4, when simply introducing semantic reconstruction to the baseline method, the performance slightly drops. By using the semantic span module, we improved the performance from 53.26% to 55.98%, demonstrating the necessity of establishing orthogonal basis vectors during semantic reconstruction.

Semantic Filtering. As shown in Table 4, directly applying semantic filtering on the baseline method harms the performance because the support features contain aliasing

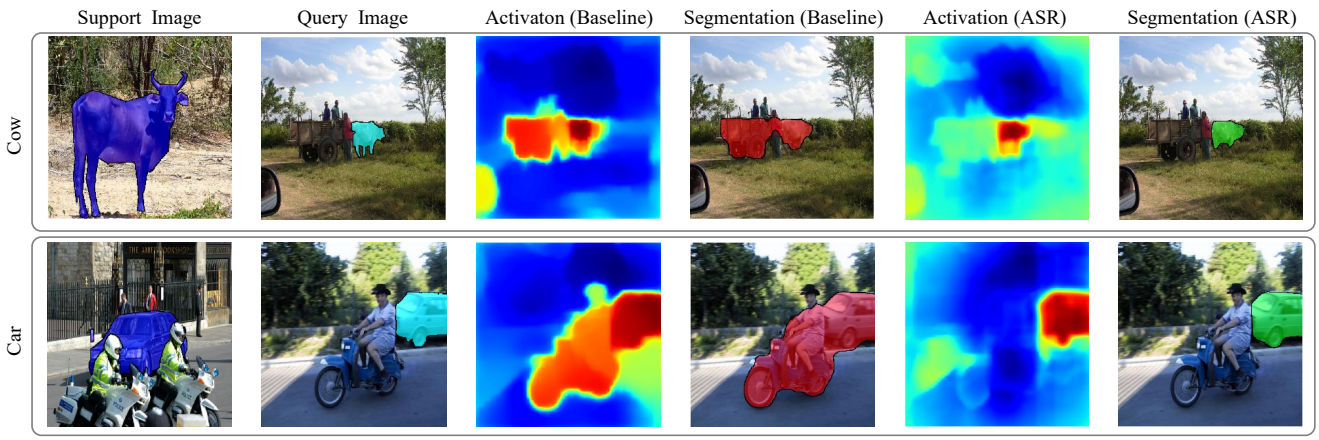


Figure 8. Semantic segmentation results. Compared with the baseline method [43], ASR (ours) reduces false positive pixels as well as activating more pixels within target object extent. (Best viewed in color)

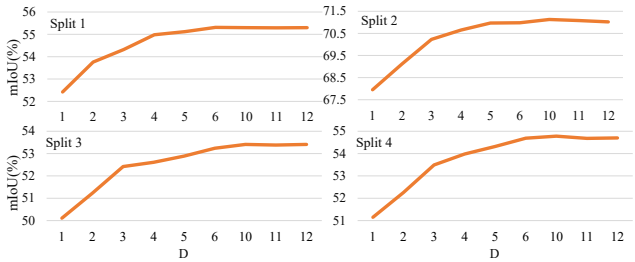


Figure 9. Performance over channel number (D) of basis vectors.

semantics. By combining all the modules, ASR improves the mIoU by 2.66% (58.64% vs. 55.98%). In Table 5, four filtering strategies are compared. The vector projection strategy defined in Section 3.4 achieves the best result. Vector projection utilizes the characteristic of vector operations to retain semantics related to the target class and suppress unrelated semantics at utmost.

Channel Number (D). The channel number (D) of features to construct basis vectors is an important parameter which affects the orthogonality of basis vectors. From Fig. 9 we can see that the performance improves with the increase of D and starts to plateau when $D = 8$, where the orthogonality of different basis vectors is sufficient for novel class reconstruction. For the MS COCO dataset D is set to 30.

Model Size and Efficiency. The model size of ASR is 36.7M, which is slightly larger than that of the baseline method [43] (36.3M) but much smaller than other methods, such as OSLSM [29] (272.6M) and FWB [25] (43.0M). With a Nvidia Tesla V100 GPU, the inference speed is 30 FPS, which is comparable with that of CANet (29 FPS).

4.5. Two-way Few-shot Segmentation

Following the settings in [24], we conduct two-way one-shot segmentation experiments on PASCAL VOC. From Tab. 6 one can see that ASR outperforms PPNet [24] with a significant margin (53.13% vs. 51.65%). Because two-way segmentation requires not only to segment targets objects but also to distinguish different classes, the model is more sensitive to semantic aliasing. Our ASR approach effectively reduces the semantic aliasing between novel classes and thereby achieves superiors segmentation performance.

Method	Pascal-5^0	Pascal-5^1	Pascal-5^2	Pascal-5^3	Mean
PPNet [24]	47.36	58.34	**52.71**	48.18	51.65
ASR (ours)	**49.35**	**60.68**	52.12	**50.38**	**53.13**

Table 6. Mean-IoU performance of 2-way 1-shot segmentation on PASCAL VOC.

5. Conclusion

We proposed Anti-aliasing Semantic Reconstruction (ASR), by converting base class features to a series of basis vectors, which span a semantic space. During training, ASR maximized the orthogonality while minimize the semantic aliasing of base classes, which facilities novel class reconstruction. During inference, ASR further suppresses interfering semantics for precise activation of target object areas. On the large-scale MS COCO dataset, ASR improved the performance of few-shot segmentation, in striking contrast with the prior approaches. As a systematic yet interpretable method for semantic representation and semantic anti-aliasing, ASR provides a fresh insight for the few-shot learning problem.

Acknowledgement. This work was supported by Natural Science Foundation of China (NSFC) under Grant 61836012, 61620106005 and 61771447, the Strategic Priority ResearchProgram of Chinese Academy of Sciences under Grant No. XDA27000000, CAAI-Huawei MindSpore Open Fund and MindSpore deep learning computing framework at www.mindspore.cn.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, pages 6230–6239, 2015. 1

[2] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 2

[3] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, and Yu-ChiangWang. A closer look at few-shot classification. In *ICLR*, 2019. 2

[4] Wen-Hsuan Chu, Yu-Jhe Li, Jing-Cheng Chang, and Yu-Chiang Frank Wang. Spot and learn: A maximum-entropy patch sampler for few-shot image classification. In *IEEE CVPR*, 2019. 2

[5] Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *IEEE CVPR*, 2021. 2

[6] Nikita Dvornik, Cordelia Schmid, and Julien Mairal. Selecting relevant features from a universal representation for few-shot classification. *CoRR*, abs/2003.09338, 2020. 2

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017. 2

[9] Siddhartha Gairola, Mayur Hemani, Ayush Chopra, and Balaji Krishnamurthy. Simpropnet: Improved similarity propagation for few-shot image segmentation. In *IJCAI*, pages 573–579, 2020. 3, 6

[10] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *IEEE CVPR*, pages 4367–4375, 2018. 2

[11] Fusheng Hao, Fengxiang He, Jun Cheng, Lei Wang, Jianzhong Cao, and Dacheng Tao. Collect and select: Semantic alignment metric learning for few-shot learning. In *IEEE ICCV*, pages 8460–8469, 2019. 2

[12] Bharath Hariharan, Pablo Arbelaez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *IEEE ICCV*, pages 991–998, 2011. 6

[13] Bharath Hariharan and Ross B. Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *IEEE ICCV*, pages 3037–3046, 2017. 2

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):386–397, 2020. 1

[15] Muhammad Abdullah Jamal and GUo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *IEEE ICCV*, pages 111719–111727, 2019. 2

[16] David C Lay, Steven R Lay, and Judi J McDonald. Linear algebra and its applications, 2016. 3, 5

[17] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting few-shot learning with adaptive margin loss. In *IEEE CVPR*, pages 12576–12584, 2020. 2

[18] Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Ji Rongrong, and Qixiang Ye. Beyond max-margin: Class margin equilibrium for few-shot object detection. In *IEEE CVPR*, June 2021. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 6

[20] Bin Liu, Yue Cao, Yutong Lin, Qi Li, Zheng Zhang, Mingsheng Long, and Han Hu. Negative margin matters: Understanding margin in few-shot classification. In *ECCV*, pages 438–455, 2020. 2

[21] B. Liu, J. Jiao, and Q. Ye. Harmonic feature activation for few-shot semantic segmentation. *IEEE Trans. Image Processing*, 30:3142–3153, 2021. 2

[22] Weide Liu, Chi Zhang, Guosheng Lin, and Fayao Liu. Crnet: Cross-reference networks for few-shot segmentation. In *IEEE CVPR*, June 2020. 2, 3, 6, 7

[23] Xiaoqian Liu, Fengyu Zhou, Jin Liu, and Lianjie Jiang. Meta-learning based prototype-relation network for few-shot classification. *Neurocomputing*, 383:224–234, 2020. 2

[24] Yongfei Liu, Xiangyi Zhang, Songyang Zhang, and Xuming He. Part-aware prototype network for few-shot semantic segmentation. In *ECCV*, pages 142–158, 2020. 3, 6, 8

[25] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *IEEE ICCV*, pages 622–631, 2019. 6, 7, 8

[26] Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alyosha A. Efros, and Sergey Levine. Conditional networks for few-shot semantic segmentation. In *ICLR Workshop*, 2018. 2, 6

[27] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 2

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241, 2015. 1

[29] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *BMVC*, 2017. 6, 8

[30] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *IEEE CVPR*, pages 1199–1208, 2018. 2

[31] Zhuotao Tian, Hengshuang Zhao, Michelle Shu, Zhicheng Yang, Ruiyu Li, and Jiaya Jia. Prior guided feature enrichment network for few-shot segmentation. abs/2008.01449, 2020. 6

[32] Oriol Vinyals, Charles Blundell, Tim Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *NeurIPS*, pages 3630–3638, 2016. 2

[33] Haochen Wang, Xudong Zhang, Yutao Hu, Y. Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *ECCV*, 2020. 2, 6, 7

[34] Kaixin Wang, JunHao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *IEEE ICCV*, pages 622–631, 2018. 2, 6, 7

[35] Yu-Xiong Wang, Ross B. Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *IEEE CVPR*, pages 7278–7286, 2018. 2

[36] Yu-Xiong Wang and Martial Hebert. Learning to learn: Model regression networks for easy small sample learning. In *ECCV*, pages 616–634, 2016. 2

[37] Boyu Yang, Chang Liu, Bohao Li, Jianbin Jiao, and Qixiang Ye. Prototype mixture models for few-shot semantic segmentation. In *ECCV*, pages 763–778, 2020. 3, 6, 7

[38] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *IEEE CVPR*, 2021. 2

[39] M. Zand, S. Doraisamy, A. Abdul Halin, and M. R. Mustaffa. Ontology-based semantic image segmentation using mixture models and multiple crfs. *IEEE Trans. Image Processing*, 25(7):3233–3248, 2016. 2

[40] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Differentiable earth mover's distance for few-shot learning, 2020. 2

[41] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. Deepemd: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *IEEE CVPR*, 2020. 2

[42] Chi Zhang, Guosheng Lin, Fayao Liu, Jiushuang Guo, Qingyao Wu, and Rui Yao. Pyramid graph networks with connection attentions for region-based one-shot semantic segmentation. In *IEEE ICCV*, 2019. 2, 6, 7

[43] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *IEEE CVPR*, pages 5217–5226, 2019. 2, 3, 5, 6, 7, 8

[44] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *IEEE CVPR*, June 2021. 2

[45] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. In *NeurIPS*, pages 147–155, 2019. 1

[46] Xiaosong Zhang, Fang Wan, Chang Liu, Xiangyang Ji, and Qixiang Ye. Learning to match anchors for visual object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–14, 2021. 1

[47] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *CoRR*, abs/1810.09091, 2018. 2, 3, 6

[48] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE CVPR*, pages 6230–6239, 2017. 1, 2

[49] Kai Zhu, Wei Zhai, and Yang Cao. Self-supervised tuning for few-shot segmentation. In *IJCAI*, pages 1019–1025, 2020. 3, 6, 7