SMILe: Leveraging Submodular Mutual Information For Robust Few-Shot Object Detection

Anay Majee[1], Ryan Sharp[2]*, and Rishabh Iyer[1]

[1] The University of Texas at Dallas, TX, USA
{anay.majee, rishabh.iyer}@utdallas.edu
[2] IGS Energy, OH, USA
rysharp@igsenergy.com

Abstract. Confusion and forgetting of object classes have been challenges of prime interest in Few-Shot Object Detection (FSOD). To overcome these pitfalls in metric learning based FSOD techniques, we introduce a novel **S**ubmodular **M**utual **I**nformation **L**earning (**SMILe** [3]) framework for loss functions which adopts combinatorial mutual information functions as learning objectives to enforce learning of well-separated feature clusters between the base and novel classes. Additionally, the joint objective in SMILe minimizes the total submodular information contained in a class leading to discriminative feature clusters. The combined effect of this joint objective demonstrates significant improvements in class confusion and forgetting in FSOD. Further we show that SMILe generalizes to several existing approaches in FSOD, improving their performance, agnostic of the backbone architecture. Experiments on popular FSOD benchmarks, PASCAL-VOC and MS-COCO show that our approach generalizes to State-of-the-Art (SoTA) approaches *improving their novel class performance by up to 5.7% (3.3 mAP points) and 5.4% (2.6 mAP points) on the 10-shot setting of VOC (split 3) and 30-shot setting of COCO datasets* respectively. Our experiments also demonstrate better retention of base class performance and *up to* 2× *faster convergence* over existing approaches agnostic of the underlying architecture.

1 Introduction

Recent advances in Deep Neural networks (DNNs) have enabled models to learn discriminative feature representations from *large-scale* image benchmarks. Unfortunately, these architectures fail to adapt to few-shot settings tasked to recognize novel objects over existing ones with few examples, closely resembling human-like perception. Although recent research has shown significant promise in few-shot image recognition [10, 12, 33, 35, 36], Few-Shot Object Detection (FSOD) remains a challenge with recent works [27, 28, 37, 38] highlighting two

* Work done as a Graduate student at The University of Texas at Dallas.
[3] Project page: https://anaymajee.me/assets/project_pages/smile.html.

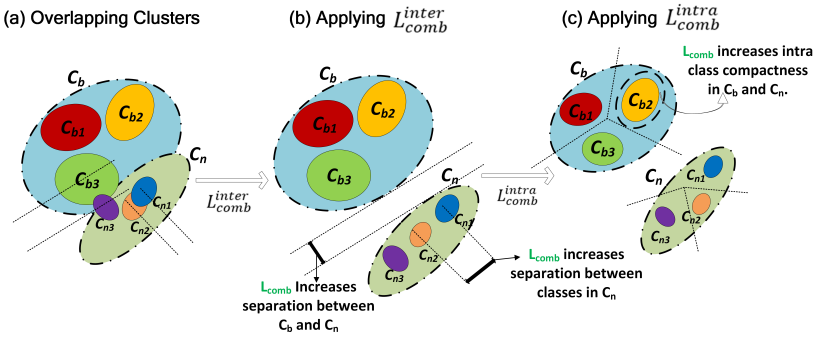

Fig. 1: Functionality of components in L_{comb} proposed in the SMILe (ours) framework, (a) L_{comb}^{inter} promotes separation between C_b and C_n while (c) L_{comb}^{intra} promotes intra-class compactness.

major challenges - *Class Confusion* and *Catastrophic Forgetting*. *Class confusion*, as highlighted in [29] manifests itself through mis-prediction of instances belonging to a newly learnt (novel) class, as one or more instances of the already learnt (base) classes. Authors in [1, 38] attribute this to the sharing of visual information between classes resulting in **increased inter-class bias** due to overlapping feature clusters as shown in Fig. 1(a). *Catastrophic forgetting* refers to the gradual degradation in the performance of already learnt classes in the quest to learn the novel ones, as shown in Fig. 2(a), seldom overfitting to rare classes [29, 38]. Further, **large feature diversity (intra-class variance) among base classes** lead to formation of non-discriminative feature clusters as shown in Fig. 1(b), aggravating the existing inter-class bias in the feature space. Unlike existing approaches (refer Sec. 2) which target either confusion or forgetting, our paper presents a unified approach to tackle both these challenges in FSOD.Although, recent approaches [1, 28, 34] attempt to tackle these challenges through contrastive learning strategies, such approaches have been limited by their capability to overcome either inter-class bias or intra-class variance [30, 38] and poor generalization to longtail settings [30] (FSOD being a extreme case).

In this paper, we introduce a combinatorial viewpoint in FSOD considering each object class $i \in [1, C]$ in the dataset $\mathcal{T}$ as a set A_i of samples, where $\mathcal{T} = \{A_1, \cdots A_C\}$, facilitating the application of combinatorial functions as learning objectives. We aim to overcome the aforementioned challenges through representation learning in the low-data regime by adopting this formulation through the **SMILe**: **S**ubmodular **M**utual **I**nformation **Le**arning framework, wherein we introduce novel, set-based combinatorial objective functions for FSOD as shown in Fig. 3. **SMILe introduces a joint objective formulation** L_{comb} **(Eq. (3)) based on two popular flavors of submodular information functions - Submodular Mutual Information [22] (SMI) and Total Submodular Information [11]** targeting the root causes of confusion and forgetting in FSOD. At first, SMILe is the first to introduce pairwise SMI functions I_f in representation learning which model the common (overlapping) information between two

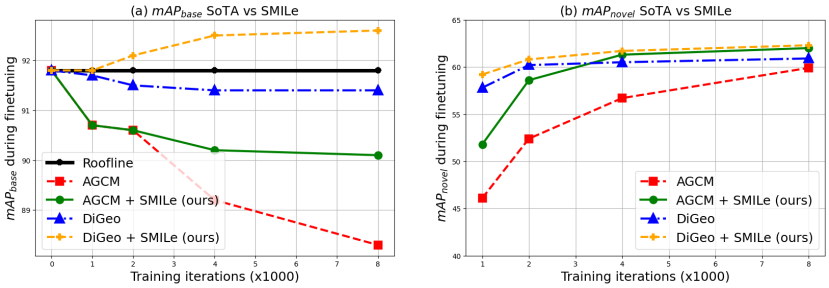

Fig. 2: Resilience to Catastrophic forgetting and faster convergence in SMILe over SoTA approaches. (a) shows that combinatorial losses in SMILe are robust to catastrophic forgetting, while (b) shows that objectives in SMILe results in faster convergence over SoTA FSOD methods (AGCM and DiGeo).

sets. Minimizing I_f through the joint objective L_{comb} reduces feature overlap between *base* and *novel* classes alleviating inter-class bias in the model towards abundantly sampled classes as shown in Fig. 1(b). We extend this property of SMI functions to the *novel* classes minimizing the inter-cluster overlap between few-shot classes, promoting learning of discriminative features from just few samples. Secondly, SMILe preserves the diversity within each class by minimizing the total submodular information contained within each set as shown in Fig. 1(c), minimizing the impact of forgetting. This formulation closely follows the observation in [30] which models cooperation [15] between instances in a set by minimizing a submodular function over a set, to preserve representative features. The unified objective L_{comb} introduced in SMILe models both these necessary properties through a weighted sum of two distinct objectives L_{comb}^{inter} and L_{comb}^{intra} as shown in Fig. 3 balancing the tradeoff between inter-cluster separation and intra-cluster compactness respectively. This allows us to introduce a family of loss functions which inherently eliminates confusion and forgetting as shown in Tab. 5. We conduct our experiments on two popular FSOD benchmarks, PASCAL-VOC [6] and MS-COCO [26] for several few-shot settings and demonstrate the following contributions of SMILe:

- **SMILe introduces a novel set-based combinatorial viewpoint in FSOD** by applying combinatorial Mutual Information based objective to discriminate between base and novel classes, in conjunction with submodular total information to minimize intra-class variance as the objective function.
- SMILe **generalizes to existing approaches in FSOD, irrespective of the underlying architecture** demonstrating up to 5.7% improvement in novel class performance (Tab. 2) over the baseline FSOD approach.
- SMILe demonstrates up to **2× faster convergence** (Fig. 2(b)) over existing SoTA approaches resulting in faster generalization to unknown object classes.
- Finally, SMILe demonstrates up to **11% and 3.5% reduction in class confusion and catastrophic forgetting** while achieving SoTA performance on popular FSOD benchmarks like PASCAL-VOC (by 5.7% on split 2, 10-shot setting) and MS-COCO (5.4% on 30-shot setting).

2 Related Work

Few-Shot Object Detection (FSOD): Classical FSOD approaches utilize finetuning [3] or distance metric learning [17] to adapt features to novel classes. Recent methods employ meta-learning techniques [16, 40, 41] with episodic training to learn class-specific features. Meta-Reweight [16] and Meta-RCNN [41] use additional feature extractors, while Add-Info [40] leverages feature differences between support and query images. Techniques like [43] enhance class-specific features through information sharing, and CME [24] aims to reduce class confusion. Attention mechanisms [7, 44] are used to identify discriminative features. However, meta-learning approaches are resource-intensive and may fail to generalize to significantly different novel classes. Metric learning strategies like FsDet [37], FSCE [34], and SRR-FSD [45] offer better generalization without additional overheads. PNPDet [42] partially addresses catastrophic forgetting and class confusion. GFSD [8] proposes a Bias-Balanced RPN to mitigate overfitting in metric learners.

Recent approaches like [18, 32] adopt weak supervision from unlabelled data or low confidence predictions in RoI pooling layers to generalize to novel classes. These methods often use abundant samples from base classes [28] to prevent catastrophic forgetting, adding computational overhead in low-shot settings. Vision transformers [4] have been adopted in FSOD through methods like imTED [27] and PDC [23], with reduced computational overhead by using pre-trained attention heads. Alternatively, DiGeo [28] and PDC [23] learn the geometry or difference in distributions of RoI proposals [13] between object classes to overcome forgetting and confusion. However, these approaches rely on contrastive learning objectives [20] that struggle to learn discriminative feature embeddings due to adoption of pairwise similarity metrics. Our work, SMILe, aims to improve the feature learning capacity of existing SoTA approaches, irrespective of their underlying architectures.

Submodular Functions and Combinatorial Objectives : Submodular functions are recognized as set functions with an inherent diminishing returns characteristic. Defined as a set function $f : 2^{\mathcal{V}} \rightarrow \mathbb{R}$ operating on a ground-set $\mathcal{V}$, a function is termed submodular if it adheres to the condition $f(X) + f(Y) \geq f(X \cup Y) + f(X \cap Y), \forall X, Y \subseteq \mathcal{V}$ [11]. These functions have garnered considerable attention in research, particularly in fields like data subset selection [22], active learning [21], and video summarization [19, 22] through their ability in modeling concepts such as diversity, relevance, set-cover and representation. A subclass of submodular functions, namely Submodular Mutual Information (SMI) functions introduced in [22] model the similarity and diversity between pairs of object classes establishing itself as a powerful tool to model inter-class bias. Recently, Majee et al. [30] introduces these set-based combinatorial functions as objectives in representation learning and demonstrates their capability in overcoming inter-class bias (by minimizing the similarity between nonidentical object classes) and intra-class variance (maximizing the similarity between instances of the same object class). However, these functions are yet to be studied in the the context of few-shot learning. We introduce novel instances of SMI based objectives in

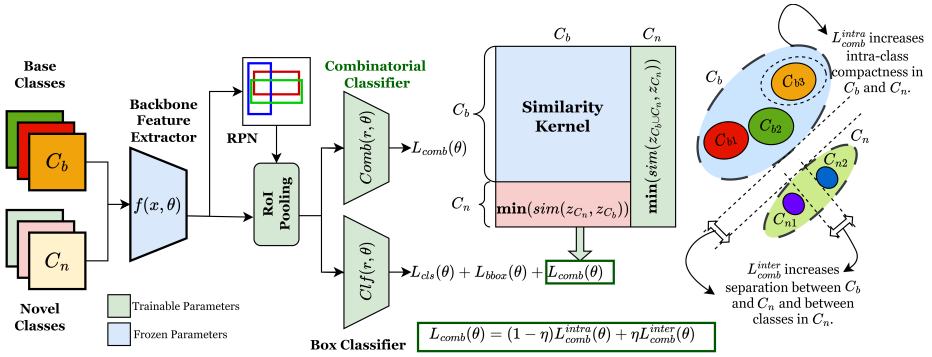

Fig. 3: Overview of our SMILe framework highlighting the application of Mutual Information function based objectives in SMILe for the fine-tuning stage of Few-Shot Object Detection.

SMILe to minimize inter-class bias between base and novel classes. To the best of our knowledge we are the first to introduce novel SMI based combinatorial objectives in conjunction with total information based combinatorial functions through SMILe in a quest to minimize confusion and forgetting in few-shot object detection.

3 Method

3.1 Problem Definition : Few-Shot Object Detection

We define a few-shot learner $h(x, \theta)$ as shown in Fig. 3 that receives input data x from base classes $C_b \in [1, |C_b|]$ and novel classes $C_n \in [1, |C_n|]$ such that $C = \{C_b \cup C_n\}$ and $\{C_b \cap C_n\} = \varnothing$. Here, θ denotes the learnable parameters. The training data can be divided into two distinct parts, base D_{base} and novel D_{novel} such that, $\mathcal{T} = \{D_{base} \cup D_{novel}\}$ and $\{D_{base} \cap D_{novel}\} = \varnothing$. **SMILe introduces a paradigm shift in FSOD by imbibing a combinatorial viewpoint**, where the base dataset, $D_{base} = [A^b_1, A^b_2, \cdots, A^b_{|C_b|}]$, containing abundant training examples from C_b base classes and the novel dataset, $D_{novel} = [A^n_1, A^n_2, \cdots, A^n_{|C_n|}]$ containing only K-shot ($|A^n_i| = K$ for $i \in [1, C_n]$) training examples from C_n novel classes. The objective of the few-shot learner $h(x, \theta)$ *is to learn discriminative representation from classes in* D_{novel} *without degradation in performance on classes in* D_{base}. Following FSCE [34] we adopt a two-stage training strategy. In the **base training** stage we train $h(x, \theta)$ on abundant samples in D_{base}, allowing the model to generalize on the domain of D_{base}. The **few-shot adaptation** stage adapts $h(x, \theta)$ to *previously unseen* K-shot data by fine-tuning on data samples from $D_{base} \cup D_{novel}$ where $|A_k| = K$ for $k \in \{C_b \cup C_n\}$. The goal of SMILe is to overcome class confusion and forgetting in FSOD resulting from elevated inter-class bias and intra-class variance as observed in [1, 29, 38]. The final model $h(x, \theta)$ obtained after two training stages is evaluated on D_{test} containing unseen data samples from both $C_b \cup C_n$.

3.2 The SMILe Framework

Adopting a combinatorial viewpoint as disclosed earlier allows us to employ submodular combinatorial functions as learning objectives to tackle confusion and forgetting in FSOD. As discussed in Sec. 2, minimizing a Submodular functions naturally models cooperation [15] while maximizing it models diversity [25] due to their inherent diminishing marginal returns property. SMILe adopts the aforementioned properties of submodular functions to *define a novel family of combinatorial objective (loss) functions $L_{comb}(\theta)$ which enforces orthogonality in the feature space when applied on Region-of-Interest (RoI) features in FSOD models*. The loss function $L_{comb}(\theta)$ can be decomposed into two major components - L_{comb}^{inter} minimizes inter-class bias between base and novel classes and L_{comb}^{intra} maximizes intra-class compactness within abundant classes.

For L_{comb}^{inter}, SMILe explores a sub-category of combinatorial functions, namely *Submodular Mutual Information* (SMI) which can be defined as $I_f(A_i, A_j) = f(A_i) + f(A_j) - f(A_i \cup A_j)$ [11, 22], and models the common information between two sets A_i and A_j, $\forall i, j \in \mathcal{T}$. Results in [11, 22] portray $I_f(A_i, A_j; \theta)$ as a measure of the degree of similarity between object classes A_i and A_j. Adopting this definition of SMI, L_{comb}^{inter} **minimizes the SMI between the base C_b and the novel C_n classes, ensuring sufficient inter-cluster separation** (by minimizing inter-class bias) as shown in Eq. (1). L_{comb}^{inter} further **minimizes the mutual information between classes in C_n**, minimizing inter-cluster overlaps between the novel classes. This is visually depicted in Fig. 1(b) and has been shown to be *effective in mitigating class confusion* in FSOD through our experiments in Sec. 4.

$$L_{comb}^{inter}(\theta) = \sum_{\substack{b \in C_b \\ n \in C_n}} I_f(A_b, A_n; \theta) + \sum_{\substack{i,j \in C_n \\ i \neq j}} I_f(A_i, A_j; \theta) = \sum_{\substack{i \in (C_b \cup C_n) \\ j \in C_n : i \neq j}} I_f(A_i, A_j; \theta) \tag{1}$$

In addition to confusion which stems from inter-class bias, SMILe aims at mitigating catastrophic forgetting [29] in FSOD which has been attributed to large intra-class variance among abundant object classes in [1,38]. In coherence to the combinatorial formulation in SMILe we achieve this through L_{comb}^{intra} which minimizes the Total Submodular Information, defined as $S_f(A_1, \cdots, A_{|C|}) = \sum_{k=1}^{|C|} f(A_k; \theta)$, over sets $A_k \in \mathcal{T}$, given a submodular function $f(A_k; \theta)$. As discussed earlier, minimizing the submodular information models cooperation which asserts that minimizing L_{comb}^{inter} promotes learning of discriminative feature clusters, penalizing abundant classes to have large feature variance in the embedding space as shown in Fig. 1(c). Although submodular functions have been studied in the field of representation learning to minimize intra-class variance in [30], but primarily differs from SMILe in modeling a longtail recognition task by minimizing the total submodular correlation, which models gain in information when new features are added to a set. The formulation of L_{comb}^{intra} has been shown in Eq. (2) where we **minimize the total submodular information within samples in each class in $C_b \cup C_n$** and our experiments in Tab. 5

show the effectiveness of L_{comb}^{intra} in boosting base class performance *asserting the mitigation of catastrophic forgetting.*

$$L_{comb}^{intra}(\theta) = \sum_{b \in C_b} f(A_b, \theta) + \sum_{n \in C_n} f(A_n, \theta) = \sum_{k \in (C_b \cup C_n)} f(A_k, \theta) \tag{2}$$

Ablating on the choice of the submodular function f and SMI functions I_f we introduce several instances of SMILe objectives as discussed in Tab. 1.

Encapsulating the aforementioned formulations of L_{comb}^{inter} and L_{comb}^{intra} in SMILe we define a joint objective $L_{comb}(\theta)$ which tackles both the challenges of confusion and forgetting. We thus define $L_{comb}(\theta)$ in Eq. (3) which is the weighted algebraic sum of L_{comb}^{inter} and L_{comb}^{intra} with the weighting factor η.

$$\begin{aligned} L_{comb}(\theta) =& (1-\eta) L_{comb}^{intra}(\theta) + \eta L_{comb}^{inter}(\theta) \\ =& \sum_{i \in C_b \cup C_n} \left[(1-\eta) f(A_i, \theta) + \eta \sum_{\substack{j \in C_n \\ i \neq j}} I_f(A_i, A_j; \theta) \right] \end{aligned} \tag{3}$$

Note, that the combinatorial objective $L_{comb}(\theta)$ is applied on output features from the RoI Pooling layers in proposal-based [31,38] architectures. To **promote adoption of SMILe agnostic of the backbone architecture** we introduce a combinatorial head $Z_{comb} = Comb(h, \theta)$ which projects the RoI features to 128-dimensional feature vectors [20], Z_{comb} on which $L_{comb}(\theta)$ is applied during the few-shot adaptation stage.

Finally, we summarize the total classification loss in SMILe as depicted in Eq. (4) as the sum over all three objectives: the classification head L_{Clf}, the box regression head L_{bbox} and the combinatorial head $L_{comb}(\theta)$. Note that the objectives proposed in SMILe apply only to $Comb(h, \theta)$ while the RoI classification and regression heads are unchanged. This follows the observations in [28,38] which warrants the boost in performance originating from learning robust feature representations for each RoI predicted by the model.

$$L_{cls}(\theta) = L_{Clf}(\theta) + L_{bbox}(\theta) + L_{comb}(\theta) \tag{4}$$

3.3 Instantiations of L_{comb}^{inter} and L_{comb}^{intra} in the SMILe Framework

Given a submodular function $f(A)$ and a Submodular Mutual Information (SMI) function $I_f(A, Q)$ over sets A and Q, we derive two instances L_{comb}^{inter} and L_{comb}^{intra} objectives in SMILe. Depending on the choice of $f(A)$ we define two instances: Facility-Location Mutual Information (SMILe-FLMI) and Graph-Cut Mutual Information (SMILe-GCMI). Inherently, both objectives adopt the cosine similarity metric $S_{ij}(\theta)$ as used in SupCon [20] which can be defined as $S_{ij}(\theta) = \frac{Z_{comb_i}^{\mathrm{T}} \cdot Z_{comb_j}}{||Z_{comb_i}|| \cdot ||Z_{comb_j}||}$ to compute similarity between sets in the learning objective. Although the similarity kernel used in SMILe is computed in a pairwise fashion, objectives defined under L_{comb} use it to only compute feature interactions

Table 1: Summary of various instantiations of SMILe highlighting the components of the combinatorial objective, L_{comb}^{inter} and L_{comb}^{intra}.

Objective	**Instances of $L_{comb}^{inter}(\theta)$**	**Instances of $L_{comb}^{intra}(\theta)$**
SMILe-GCMI (ours)	$\sum_{\substack{k\in(C_b\cup C_n)\\ l\in C_n:k\neq l}} 2\lambda\sum_{i\in A_k}\sum_{j\in A_l} S_{ij}(\theta)$	$\sum_{k\in C_b\cup C_n}\sum_{i\in A_k}\sum_{j\in\mathcal{T}\setminus A_k} S_{ij}(\theta) - \lambda\sum_{i,j\in A_k} S_{ij}(\theta)$
SMILe-FLMI (ours)	$\sum_{\substack{k\in(C_b\cup C_n)\\ l\in C_n:k\neq l}} \sum_{i\in A_k}\max_{j\in A_l} S_{ij}(\theta) + \lambda\sum_{i\in A_l}\max_{j\in A_k} S_{ij}(\theta)$	$\sum_{k\in C_b\cup C_n}\sum_{i\in\mathcal{T}\setminus A_k}\max_{j\in A_k} S_{ij}(\theta)$

between samples, **differing from existing approaches in aggregation of pairwise similarities to compute total information and mutual information** over classes in $\mathcal{T}$.

SMILe-FLMI based objective is derived from the Facility-Location Mutual Information (FLMI) [22] function, expressed as $I_f(Q,A) = \sum_{i\in Q}\max_{j\in A} S_{ij}(\theta) + \lambda\sum_{i\in A}\max_{j\in Q} S_{ij}(\theta)$ and minimizes the maximum similarity (most similar) between sets Q and A. Given the facility-location (FL) submodular function $f(A,\theta) = \sum_{i\in\mathcal{T}}\max_{j\in A} S_{ij}(\theta)$ over the set A, we can derive $L_{comb}^{inter}(\theta)$ and $L_{comb}^{intra}(\theta)$ shown in Eq. (5) as the *SMILe-FLMI* objective. Note that $L_{comb}^{inter}(\theta)$ is applied between object classes in $C_b \cup C_n$ and C_n while $L_{comb}^{intra}(\theta)$ is applied over all classes in $C_b \cup C_n$.

$$
\begin{aligned}
L_{comb}^{inter}(\theta) &= \sum_{\substack{k\in(C_b\cup C_n)\\ l\in C_n:k\neq l}}\sum_{i\in A_k}\max_{j\in A_l} S_{ij}(\theta) + \lambda\sum_{i\in A_l}\max_{j\in A_k} S_{ij}(\theta),\\
L_{comb}^{intra}(\theta) &= \sum_{k\in C_b\cup C_n}\sum_{i\in\mathcal{T}\setminus A_k}\max_{j\in A_k} S_{ij}(\theta)
\end{aligned}
\tag{5}
$$

Minimizing the L_{comb}^{inter} objective function ensures that the sets $A_l \in C_n$ and $A_k \in C_b \cup C_n$ are disjoint by minimizing the similarity between features in A_k and the *hardest negative* ($\sum_{i\in A_k}\max_{j\in A_l} S_{ij}(\theta)$ for $k \in C_b \cup C_n$ and $l \in C_n$) feature vectors in A_l. Further, L_{comb}^{inter} enforces sufficient separation between the novel classes themselves to promote learning of disjoint feature clusters even with few-shot data overcoming confusion. Additionally, L_{comb}^{intra} minimizes the total information contained in each set $A_k \in (C_b \cup C_n)$. This objective retains discriminative feature information from each class in $\mathcal{T}$ reducing the impact of forgetting.

SMILe-GCMI based objective described in Eq. (6) minimizes the pairwise similarity of feature vectors between a positive set $A_k \in C_b \cup C_n$ and the sets in $A_l \in C_n$ while maximizing the similarity between features in each set $A_k \in C_b \cup C_n$. Given two sets Q and A, [22] defines the Graph-Cut SMI to be $I_f(Q,A) = 2\lambda\sum_{i\in Q}\sum_{j\in A} S_{ij}(\theta)$, where the Graph-Cut function over a set A is given by $f(A,\theta) = \sum_{i\in A}\sum_{j\in\mathcal{T}\setminus A_k} S_{ij}(\theta) - \lambda\sum_{i,j\in A} S_{ij}(\theta)$. Given the Graph-Cut and the Graph-Cut SMI functions, we derive $L_{comb}^{inter}(\theta)$ and $L_{comb}^{intra}(\theta)$

shown in Eq. (6) as the *SMILe-GCMI* objective. Similar to SMILe-FLMI, the $L^{inter}_{comb}(\theta)$ is applied between object classes in $C_b \cup C_n$ and C_n while $L^{intra}_{comb}(\theta)$ is applied over all classes in $C_b \cup C_n$.

$$\begin{aligned} L^{inter}_{comb}(\theta) &= \sum_{\substack{k \in (C_b \cup C_n) \\ l \in C_n : k \neq l}} 2\lambda \sum_{i \in A_k} \sum_{j \in A_l} S_{ij}(\theta), \\ L^{intra}_{comb}(\theta) &= \sum_{k \in C_b \cup C_n} \sum_{i \in A_k} \sum_{j \in \mathcal{T} \setminus A_k} S_{ij}(\theta) - \lambda \sum_{i,j \in A_k} S_{ij}(\theta) \end{aligned} \tag{6}$$

Although objectives in SMILe-FLMI and SMILe-GCMI are tasked with similar functions, the L^{inter}_{comb} in SMILe-GCMI minimizes the pairwise similarity between sets in $C_b \cup C_n$ and C_n rather than the most similar set in SMILe-FLMI. Further, the L^{intra}_{comb} in SMILe-GCMI scales linearly with size of A_k as described in [30]. This does not allow the model to substantially improve performance on learning discriminative feature representations for classes in both C_b and C_n as the $|A_k| = K$ (number of shots) thus failing to outperform the model trained using SMILe-FLMI.

The detailed derivations of the aforementioned instances are included in the Supplementary material. Our experiments in Sec. 4.4 elucidates the fact that SMILe-FLMI is a better choice to overcome forgetting and confusion in FSOD.

4 Experiments

We evaluate models in SMILe by adopting standard evaluation criterion in FSOD [16, 37] and report the Mean Average Precision (mAP) at 50% Intersection Over Union (IoU) for all our experiments.

4.1 Experimental Setup

Datasets We evaluate our proposed SMILe approach on two few-shot object detection datasets - and PASCAL-VOC [5] and MS-COCO [26] datasets.

PASCAL-VOC [5] dataset consists of 20 classes, out of which 15 are considered as base and 5 as novel classes. The novel classes are chosen at random giving rise to three data splits namely, split-1 (*bird*, *bus*, *cow*, *motorbike*, *sofa*), split-2 (*aeroplane*, *bottle*, *cow*, *horse*, *sofa*) and split-3 (*boat*, *cat*, *motorbike*, *sheep*, *sofa*). Following previous works [16], we use the combined VOC 07+12 datasets for training and evaluate our models on the complete validation set of VOC 2007 for 1, 5, and 10 shot settings.

MS-COCO [26] dataset consists of 80 classes, out of which 60 are considered as base and 20 as novel classes. Following existing approaches in FSOD [41] we randomly select 5k samples from $(D_{base} \cup D_{novel})$ to use as the validation set while the remaining samples are used to generate random 10 and 30-shot splits for training of the MS-COCO 2014 dataset. The key difference between VOC and COCO is the large intra-class variance and class-imbalance in COCO.

Table 2: Quantitative analysis on PASCAL-VOC dataset: Few-shot object detection performance (mAP_{novel}) on novel class splits of PASCAL-VOC dataset. We tabulate results for K=1, 5, 10 shots from various SoTA techniques in FSOD. * indicates that the results are averaged over 10 random seeds. † indicates a meta-learning strategy (N-way, K-shot training).

Method	Learner Type	Backbone	Split 1			Split 2			Split 3		
			K=1	5	10	1	5	10	1	5	10
† Meta-RCNN [41]	Meta	FRCN-R101	19.9	45.7	51.5	10.4	34.8	45.4	14.3	41.2	48.1
†Meta-Reweight [16]	Meta	YOLO V2	14.8	33.9	47.2	15.7	30.1	40.5	21.3	42.8	45.9
†MetaDet [38]	Meta	FRCN-R101	18.9	36.8	49.6	21.8	31.7	43.0	20.6	43.9	44.1
†Add-Info [40]	Meta	FRCN-R101	24.2	49.1	57.4	21.6	37.0	45.7	21.2	43.8	49.6
†CME [24]	Meta	YOLO V2	17.8	44.8	47.5	12.7	33.7	40.0	15.7	44.9	48.8
PNPDet [42]	Metric	DLA-34	18.2	-	41.0	16.6	-	36.4	18.9	-	36.2
FsDet w/ FC [37]	Metric	FRCN-R101	36.8	55.7	57.0	18.2	35.5	39.0	27.7	48.7	50.2
FsDet w/ cos [37]	Metric	FRCN-R101	39.8	55.7	56.0	23.5	35.1	39.1	30.8	49.5	49.8
Retentive-RCNN [9]	Metric	FRCN-R101	40.1	53.7	56.1	21.7	37.0	40.3	30.2	49.7	50.1
FSCE [34]	Metric	FRCN-R101	41.0	57.4	57.8	27.3	44.4	49.8	40.1	53.2	57.7
FSCE + SMILe (ours)	**Comb.**	**FRCN-R101**	**41.2**	**57.9**	**61.1**	**29.2**	**44.6**	**50.5**	**41.3**	**55.6**	**59.0**
AGCM [1]	Metric	FRCN-R101	40.3	58.5	59.9	27.5	49.3	50.6	42.1	54.2	58.2
AGCM + SMILe (ours)	**Comb.**	**FRCN-R101**	**40.9**	**59.7**	**62.0**	**31.9**	**49.5**	**52.3**	**42.6**	**56.4**	**61.4**
DiGeo [28]	Metric	FRCN-R101	36.0	54.1	60.9	20.7	42.8	47.1	27.5	47.3	52.9
DiGeo + SMILe(ours)	**Comb.**	**FRCN-R101**	**36.1**	**56.6**	**62.3**	**26.5**	**44.1**	**47.3**	**33.1**	**51.9**	**56.4**
imTED [27]	Metric	ViT-B	31.9	71.9	77.0	22.7	52.2	57.7	12.6	69.6	72.8
imTED + PDC [23]	Metric	ViT-B	**36.6**	73.1	77.1	15.5	51.8	56.0	**18.9**	67.9	72.8
PDC + SMILe (ours)	**Comb.**	**ViT-B**	**36.6**	**75.2**	**77.9**	**27.1**	**52.7**	**58.3**	15.1	**70.0**	**74.7**

Implementation Details The SMILe framework adopts a architecture agnostic approach and adopt several backbones including Faster-RCNN [31] and ViT [27]. For VOC, the input batch size to the network is set to 16 and 2 in the base training and few-shot adaptation stages for Faster-RCNN and ViT based approaches. The input resolution is set to 764 x 1333 pixels for data splits in COCO, while it is set to 800 x 600 pixels for PASCAL-VOC. The hyper-parameters used in the formulation of SMILe, namely η and similarity kernel S, are chosen through ablation experiments described in Sec. 4.4. Results from existing methods are a reproduction of the algorithm from publicly available codebases. All our experiments are performed on 4 NVIDIA GTX 1080 Ti GPUs with additional details in the supplementary material and code released at `https://github.com/amajee11us/SMILe-FSOD.git`.

4.2 Results on Few-Shot PASCAL VOC Dataset

Table 2 records the results obtained from our SMILe framework on novel splits of the PASCAL-VOC dataset and contrasts it against SoTA FSOD techniques. We adopt four SoTA approaches FSCE [34], AGCM [1], DiGeo [28] and imTED [27], covering several backbone architectures Faster-RCNN + FPN (FSCE, AGCM and DiGeo) alongside ViT (imTED, PDC) and introducing SMILe ($M+SMILe$) approach into existing architectures M. For Faster-RCNN based architectures (FSCE) we show a maximum of 5.7% (3.3 mAP points) improvement while for FPN based arcitectures (AGCM and DiGeo) we show a 3.5% (2.1 mAP points

for AGCM+SMILe) improvement. It is interesting to note that unlike FSCE and AGCM, DiGeo uses abundant samples from C_b alongside few-shot samples in C_n (with upsampling) during finetuning introducing a large inter-class bias. SMILe outperforms DiGeo by up to 2.3 mAP points (split 2, 10-shot) showing the resilience of SMILe towards imbalance, thus overcoming confusion in FSOD. Additionally, for recently introduced transformer based architectures (imTED + PDC [23]) SMILe outperforms the existing SoTA with a maximum improvement of 4.9 mAP points (split 2, 5-shot) thus establishing SMILe as the SoTA on few-shot splits of VOC. Note, that the choice of objective functions L_{comb}^{inter} and L_{comb}^{intra} for this experiment has been determined to be SMILe+FLMI through an ablation on the different instances in SMILe as described in Sec. 4.4. Finally, Fig. 2(b) shows that **introduction of SMILe framework to existing SoTA approaches leads to rapid convergence on the novel classes up to 2x over existing SoTA**. This is significant for mission critical tasks like autonomous driving where the model is required to rapidly learn novel objects to reduce turn-around time.

4.3 Results on Few-Shot MS-COCO Dataset

Similar to the results in PASCAL VOC we demonstrate the results of our SMILe framework on MS-COCO dataset. In contrast to VOC, COCO presents an extremely imbalanced setting with a long-tail distribution within D_{base} itself making it really hard for FSOD approaches to achieve SoTA through primitive objective functions. Following the ablation experiments in Sec. 4.4 we adopt the SMILe+FLMI objective (best performing) to conduct the experiments on 20 few-shot classes of MS-COCO dataset. We show that SMILe generalizes existing SoTA approach (imTED + PDC) for COCO dataset by 5.4% (2.6 mAP points, 30-shot setting). This further establishes the generalizability of our approach over varying data distributions (VOC and COCO) while achieving SoTA in FSOD tasks.

4.4 Ablation Study

We conduct ablation experiments on the 10-shot split of VOC (split 1) with hyper-parameters $\eta = 0.5$, cosine similarity metric and $\lambda = 1.0$. Ablations for hyper-parameters are detailed in the supplementary material.

Components of SMILe Instantiations in SMILe consists of two main components - L_{comb}^{inter} and L_{comb}^{intra}. We consider three baselines FSCE, AGCM and DiGeo which follow the Faster-RCNN/FPN backbone for this experiment. First, we introduce L_{comb}^{intra} by adopting the FL based objective as determined through ablation experiments below. This objective models intra-class variance and ensures reduction in intra-class variance characterized by boost in base class performance. Secondly, following the SMILe-FLMI formulation in Eq. (5) we introduce L_{comb}^{inter} during the few-shot adaptation stage. Applying this objective minimizes

Table 3: Performance of SMILe on MS COCO dataset : Our SMILe objectives demonstrate better generalizability while outperforming SoTA FSOD approaches on novel class performance mAP_{50} (novel).

Method	mAP	mAP_{50}	mAP_{75}	mAP	mAP_{50}	mAP_{75}
	10-shot			**30-shot**		
Meta-Reweight [16]	5.6	12.3	4.6	9.1	19.0	7.6
Meta-RCNN [41]	8.7	19.1	6.6	12.4	25.3	10.8
TFA w/cos [37]	10.0	-	9.3	13.7	-	13.4
Add-Info [40]	12.5	27.3	9.8	14.7	30.6	12.2
MPSR [39]	9.8	17.9	9.7	14.1	25.4	14.2
FSCE [34]	11.9	-	10.5	16.4	-	16.2
FADI [2]	12.2	22.7	11.9	16.1	29.1	15.8
CME [24]	15.1	24.6	16.4	16.2	-	-
FCT [14]	17.1	30.2	17.0	21.4	35.5	22.1
imTED-B [27]	22.5	36.6	23.7	30.2	47.4	32.5
imTED-B+PDC [23]	23.4	38.1	24.5	30.8	47.3	33.5
PDC + SMILe (ours)	**25.8**	**40.1**	**26.1**	**31.0**	**49.9**	**33.6**

Table 4: Ablation on various components of the proposed SMILe approach.

Method	Baseline	$f(A_i)$ (L_{comb}^{intra})	$I_f(A_i, A_j)$ (L_{comb}^{inter})	mAP_{base}	mAP_{novel}
FsDet w/ cos	-	-	-	23.6	39.8
FSCE	✓			86.1	57.8
	✓	✓		89.6	60.1
	✓		✓	88.3	61.0
	✓	✓	✓	**89.8**	**61.1**
AGCM	✓			87.6	58.0
	✓	✓		88.6	61.3
	✓		✓	88.9	61.8
	✓	✓	✓	**89.3**	**61.8**
DiGeo	✓			90.5	60.9
	✓	✓		92.3	61.7
	✓		✓	91.4	62.0
	✓	✓	✓	**92.6**	**62.3**

the inter-class bias between $C_b \cup C_n$ and C_n, thus improving novel class performance significantly. Nevertheless, we see a slight drop in base class performance due to forgetting prevalent in FSOD tasks. Finally, we combine both instantiations in SMILe into one single objective as in Eq. (3) with $\eta = 0.5$ and show that SMILe improves both base and novel class performance emerging as the best choice for FSOD. To demonstrate generalization, we perform this experiment on several SoTA approaches as baseline and show that the results discussed in aforementioned section holds. We summarize all the results in Tab. 4.

Choice of Combinatorial functions in L_{comb} SMILe introduces several instances of L_{comb}^{intra} and L_{comb}^{inter}. To clearly understand the contributions of each of these instances we conduct experiments tabulated in Tab. 5 and determine the

Table 5: Ablation on the choice of Submodular Information function I_f and Submodular Information function f for L_{comb} in SMILe.

Model	$f(A,\theta)$	$I_f(A_i, A_j)$	mAP_{base}	mAP_{novel}
DiGeo [28]	-	-	87.9	57.4
	GC	GCMI	92.6	60.9
	FL	**FLMI**	**93.1**	**62.3**
FSCE [34]	-	-	86.1	57.8
	GC	GCMI	87.4	60.3
	FL	**FLMI**	**89.8**	**61.1**

best performing formulation which generalizes to existing FSOD architectures. Unlike other ablation experiments in Sec. 4.4, we conduct our experiments on DiGeo which introduces an extremely imbalanced scenario by using abundant samples in D_{base}. We conclude that SMILe-FLMI which considers the Facility-Location based objective as L_{comb}^{intra} and Facility-Location Mutual Information based objective as L_{comb}^{inter} as the best performing instantiation. This result follows the formulation in Eq. (5) where FL naturally models intra-class compactness in class-imbalanced settings [30] while FLMI penalizes the classes in C_b to learn overlapping feature representations with the classes in C_n. **We use SMILe-FLMI for all benchmark experiments in Sec. 4.2 and Sec. 4.3**.

Robustness to Catastrophic Forgetting One of the most significant challenges in FSOD is the elimination of catastrophic forgetting which manifests as the degradation in the performance of classes in C_b while learning classes in C_n. This primarily occurs due to the lack of discriminative feature representations from instances in D_{base} during the few-shot adaptation (stage 2) stage. We plot the change in base class performance as the training progresses in existing SoTA methods AGCM and DiGeo against number of training iterations in Fig. 2(a). At first, we contrast the change in base class performance mAP_{base} between AGCM and AGCM+SMILe and observe that AGCM overfits on the few-shot samples in D_{base} reducing the performance on C_{base} as the training progresses. AGCM + SMILe on the other hand better retains the performance on base classes with $\sim$3.5% better retention in base class performance. Interestingly, DiGeo is able to retain most of the base class performance with a very small degradation over the roofline (a model trained with only the base classes until convergence). **Our Digeo+SMILe approach outperforms Digeo by demonstrating base class performance even higher than the roofline** establishing the supremacy of L_{comb} in overcoming inter-class bias and intra-class variance resulting in robustness against catastrophic forgetting.

Overcoming Class Confusion Figure 4 highlights the supremacy of the proposed SMILe framework in mitigating class confusion through confusion matrix plots. We compare the confusion between classes in $C_b \cup C_n$ of SoTA approaches AGCM and DiGeo before and after introduction of combinatorial objectives in SMILe. Although both approaches use K-shot examples for classes in C_n, Di-

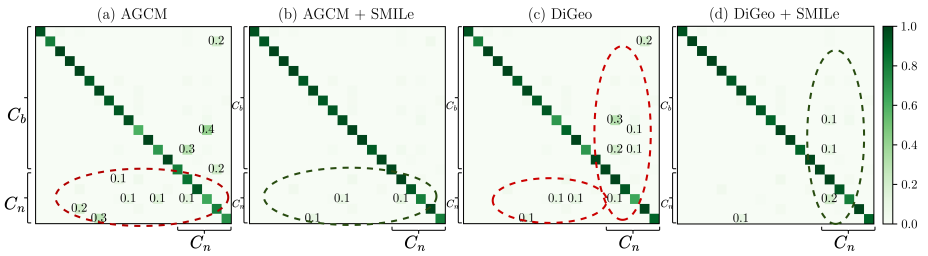

Fig. 4: Ablation on Overcoming Class Confusion in SMILe. (a,b) SMILe demonstrates 11% lower confusion over AGCM and (c,d) 4% lower confusion over DiGeo. Only significant numbers are highlighted. Best viewed in 200% zoom.

Geo differs from AGCM by adopting an upsampling strategy which allows the utilization of abundant examples in C_b while upsampling the instances in C_n. This injects different degrees of inter-class biases for models trained by adopting AGCM and DiGeo which has been demonstrated as the primary reason for confusion in previous work [29]. At first, we observe from Fig. 4 that by adopting the upsampling based strategy, DiGeo achieves very low confusion between already learnt base classes, leading to significantly lower confusion (5% among C_b and C_n). Further, confusion matrix plots in Fig. 4 show that AGCM+SMILe demonstrates 11% lower confusion than AGCM and DiGeo+SMILe shows 4% lower confusion and DiGeo. This proves the efficacy of combinatorial objectives (L^{inter}_{comb}) in mitigating inter-class bias, thereby reducing confusion between classes.

5 Conclusion

In this work, we have presented a novel approach to Few-Shot Object Detection (FSOD) by introducing a combinatorial viewpoint through the **SMILe** framework. By leveraging the properties of set-based combinatorial functions, SMILe aims to address the challenges of class confusion and catastrophic forgetting, which are prevalent in FSOD tasks. Our approach incorporates Submodular Mutual Information (SMI) and Submodular Information Measures (SIM) to penalize overlapping features between base and novel classes and to ensure the formation of compact feature clusters, respectively. The experimental results on PASCAL-VOC and MS-COCO benchmarks demonstrate the effectiveness of SMILe, showing significant improvements in novel class performance, faster convergence, and a reduction in class confusion and catastrophic forgetting. Overall, SMILe offers a promising direction for advancing the state-of-the-art in FSOD by providing a generalized framework that is adaptable to various underlying architectures and capable of handling the complexities associated with few-shot learning in object detection.

Acknowledgements

We gratefully thank anonymous reviewers for their valuable comments. This work is supported by the National Science Foundation under Grant Numbers IIS-2106937, a gift from Google Research, an Amazon Research Award, and the Adobe Data Science Research award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, Google or Adobe.

References

1. Agarwal, A., Majee, A., Subramanian, A., Arora, C.: Attention guided cosine margin to overcome class-imbalance in few-shot road object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops. pp. 221–230 (2022)
2. Cao, Y., Wang, J., Jin, Y., Wu, T., Chen, K., Liu, Z., Lin, D.: Few-shot object detection via association and discrimination. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021)
3. Chen, H., Wang, Y., Wang, G., Qiao, Y.: LSTD: A Low-Shot Transfer Detector For Object Detection. In: AAAI. pp. 2836–2843 (2018)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
5. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. IJCV pp. 303–338 (2010)
6. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision **88**, 303–338 (06 2010)
7. Fan, Q., Zhuo, W., Tang, C.K., Tai, Y.W.: Few-Shot Object Detection With Attention-RPN And Multi-Relation Detector. In: CVPR (2020)
8. Fan, Z., Ma, Y., Li, Z., Sun, J.: Generalized Few-Shot Object Detection Without Forgetting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4527–4536 (June 2021)
9. Fan, Z., Ma, Y., Li, Z., Sun, J.: Generalized few-shot object detection without forgetting (2021)
10. Finn, C., Abbeel, P., Levine, S.: Model-Agnostic Meta-Learning For Fast Adaptation Of Deep Networks. In: ICML (2017)
11. Fujishige, S.: Submodular Functions and Optimization, vol. 58. Elsevier (2005)
12. Gidaris, S., Komodakis, N.: Dynamic Few-Shot Visual Learning Without Forgetting. In: CVPR (2018)
13. Girshick, R.B.: Fast R-CNN. ICCV (2015)
14. Han, G., Ma, J., Huang, S., Chen, L., Chang, S.F.: Few-shot object detection with fully cross-transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5321–5330 (2022)
15. Jegelka, S., Bilmes, J.: Submodularity beyond submodular energies: Coupling edges in graph cuts. In: CVPR 2011 (2011)

16. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot Object Detection Via Feature Reweighting. In: ICCV (2019)
17. Karlinsky, L., Shtok, J., Harary, S., Schwartz, E., Aides, A., Feris, R., Giryes, R., Bronstein, A.M.: RepMet: Representative-Based Metric Learning For Classification And Few-Shot Object Detection. In: CVPR (2019)
18. Kaul, P., Xie, W., Zisserman, A.: Label, Verify, Correct: A Simple Few-Shot Object Detection Method. In: IEEE Conference on Computer Vision and Pattern Recognition (2022)
19. Kaushal, V., Iyer, R., Doctor, K., Sahoo, A., Dubal, P., Kothawade, S., Mahadev, R., Dargan, K., Ramakrishnan, G.: Demystifying multi-faceted video summarization: Tradeoff between diversity, representation, coverage and importance. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 452–461 (2019)
20. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems (2020)
21. Kothawade, S., Ghosh, S., Shekhar, S., Xiang, Y., Iyer, R.K.: Talisman: Targeted active learning for object detection with rare classes and slices using submodular mutual information. In: Computer Vision - ECCV 2022 - 17th European Conference (2022)
22. Kothawade, S., Kaushal, V., Ramakrishnan, G., Bilmes, J.A., Iyer, R.K.: PRISM: A rich class of parameterized submodular information measures for guided data subset selection. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI. pp. 10238–10246 (2022)
23. Li, B., Liu, C., Shi, M., Chen, X., Ji, X., Ye, Q.: Proposal distribution calibration for few-shot object detection. IEEE transactions on neural networks and learning systems (2022)
24. Li, B., Yang, B., Liu, C., Liu, F., Ji, R., Ye, Q.: Beyond Max-Margin: Class Margin Equilibrium For Few-Shot Object Detection. In: CVPR (June 2021)
25. Lin, H., Bilmes, J.: A class of submodular functions for document summarization. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (2011)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision – ECCV 2014. pp. 740–755. Springer International Publishing, Cham (2014)
27. Liu, F., Zhang, X., Peng, Z., Guo, Z., Wan, F., Ji, X., Ye, Q.: Integrally Migrating Pre-trained Transformer Encoder-decoders for Visual Object Detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6825–6834 (2023)
28. Ma, J., Niu, Y., Xu, J., Huang, S., Han, G., Chang, S.F.: Digeo: Discriminative geometry-aware learning for generalized few-shot object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2023)
29. Majee, A., Agrawal, K., Subramanian, A.: Few-Shot Learning For Road Object Detection. In: AAAI Workshop on Meta-Learning and MetaDL Challenge. vol. 140, pp. 115–126 (2021)
30. Majee, A., Kothawade, S.N., Killamsetty, K., Iyer, R.K.: SCoRe: Submodular Combinatorial Representation Learning. In: Forty-first International Conference on Machine Learning (ICML) (2024)

31. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection With Region Proposal Networks. IEEE Trans. on Pattern Analysis and Machine Intelligence (2015)
32. Shangguan, Z., Rostami, M.: Identification of novel classes for improving few-shot object detection (2023)
33. Snell, J., Swersky, K., Zemel, R.: Prototypical Networks For Few-shot Learning. In: NeurIPS. pp. 4077–4087 (2017)
34. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: FSCE: Few-Shot Object Detection Via Contrastive Proposal Encoding. In: CVPR (June 2021)
35. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning To Compare: Relation Network For Few-Shot Learning. In: CVPR (June 2018)
36. Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., Wierstra, D.: Matching Networks For One Shot Learning. In: NeurIPS (2016)
37. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly Simple Few-Shot Object Detection. In: ICML (2020)
38. Wang, Y.X., Ramanan, D., Hebert, M.: Meta-Learning To Detect Rare Objects. In: ICCV (2019)
39. Wu, J., Liu, S., Huang, D., Wang, Y.: Multi-scale positive sample refinement for few-shot object detection. In: European Conference on Computer Vision (2020)
40. Xiao, Y., Marlet, R.: Few-Shot Object Detection And Viewpoint Estimation For Objects In The Wild. In: ECCV (2020)
41. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta R-CNN: Towards General Solver For Instance-Level Low-Shot Learning. In: CVPR. pp. 9577–9586 (2019)
42. Zhang, G., Cui, K., Wu, R., Lu, S., Tian, Y.: PNPDet: Efficient Few-Shot Detection Without Forgetting Via Plug-And-Play Sub-Networks. In: WACV. pp. 3823–3832 (2021)
43. Zhang, L., Zhou, S., Guan, J., Zhang, J.: Accurate Few-Shot Object Detection With Support-Query Mutual Guidance And Hybrid Loss. In: CVPR (June 2021)
44. Zhang, S., Luo, D., Wang, L., Koniusz, P.: Few-Shot Object Detection By Second-order Pooling. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (November 2020)
45. Zhu, C., Chen, F., Ahmed, U., Shen, Z., Savvides, M.: Semantic Relation Reasoning For Shot-Stable Few-Shot Object Detection. In: CVPR (June 2021)