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

Anay Majee[1], Ryan Sharp[1]*, 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

1 Notation

Following the problem definition in Sec. 3.1 we introduce several important notations in Table 1 that are used throughout the paper.

Table 1: Collection of notations used in the paper.

Symbol	Description
$\mathcal{T}$	The training Set. $\|\mathcal{T}\|$ denotes the size of the training set.
$h(x, \theta)$	Feature extractor without the box classifier and regressor.
$Clf(., \theta)$	Multi-Layer Perceptron as classifier and regression head (as in Faster-RCNN).
$Comb(., \theta)$	Multi-Layer Perceptron as Combinatorial Classifier head.
θ	Parameters of the feature extractor.
$S_{ij}(\theta)$	Similarity between images $i, j \in \mathcal{T}$.
C_b	Classes indexed in the base dataset.
C_n	Classes indexed in the novel dataset.
C	All classes in the input dataset $\mathcal{T}$ represented as $C_b \cup C_n$.
A_k	Target set containing feature representation from a single class $k \in C$.
$f(A)$	Submodular Information function over a set A.
$I_f(A, Q)$	Mutual information function between set A and Q.
$L(\theta)$	Loss value computed over all classes $i \in C$.
L_{comb}	Combiatorial Objectives in SMILe.

2 Implementation Details

As discussed in the main paper the SMILe framework proposes an architecture agnostic approach and adopts several backbones - Faster-RCNN [8] and ViT [4, 6]. We conduct experiments on PASCAL-VOC [3] and COCO [5] datasets. For VOC, the input batch size to the network (both Faster-RCNN and ViT based approaches) is set to 16 and 2 in the base training and few-shot adaptation stages

* Work done as a Graduate student at The University of Texas at Dallas.

for Faster-RCNN and ViT based approaches. Our experiments in Tab. 2 of the main paper applies the combinatorial formulation in SMILe to four different architectures - FSCE [9], AGCM [1], DiGeo [7] and imTED+PDC [4].

For FSCE and AGCM we train the model for a maximum of 12k iterations and 6k iterations respectively with an initial learning rate of 0.01 with a batch size of 16 for VOC and 8 for COCO datasets. For DiGeo and DiGeo+SMILe we train the model for 15k steps with 200 warmup steps with a batch size of 8 and an initial learning rate of 0.05 for both datasets. The codebase for AGCM + SMILe and FSCE + SMILe has been released at `https://github.com/amajee11us/SMILe-FSOD.git`. For the DiGeo + SMILe architecture we follow the authors in [7] and introduce $Comb(h, \theta)$ in the *distill* stage of the training process. Following the authors in [7] we use abundant samples of the base classes and K-shot (few-shot) samples of the novel classes and use the same set of hyper-parameters as released in our codebase at `https://github.com/amajee11us/SMILe-FSOD/tree/digeo`.

Due to adoption of ViT [2] based architecture in imTED + PDC and imTED + PDC + SMILe architectures, we train the model with a batch size of 2 (as used in [4]) with an initial learning rate of 1e-4 for a total of 108 epochs with a step learning rate scheduler. We release the code for training and inferencing on the PDC + SMILe is released at `https://github.com/amajee11us/SMILe-FSOD/tree/pdc_SMILe`.

The $Comb(h, \theta)$ architecture is applied only during the few-shot adaption stage (across architectures) of model training and 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. For all architecture variants we adopt the Stronger Baseline introduced in FSCE [9] with a trainable Region Proposal Network (RPN) and RoI Pooling layer alongside increasing the number of RoI proposals to 2048 (double the number as compared to [10]). The additional RoI proposal features help capture the low confidence novel classes in the initial training iterations leading to faster convergence. Additionally we introduce two hyper-parameters in the formulation of SMILe, namely η and similarity kernel S, are chosen through ablation experiments described in Sec. 3. Following existing research [1, 9] we report the novel class performance for 1, 5, 10 shot settings for VOC and 10, 30 settings for COCO averaged over 10 distinct seeds[3]. Results from existing methods are a reproduction of the algorithm from publicly available codebases.

3 Ablation : Hyper-Parameters in SMILe

We perform ablation on various hyper-parameters introduced in SMILe and derive their values which lead to the best possible base and novel class performance in the few-shot adaptation stage. For all our experiments we consider the AGCM [1] architecture as the baseline and train and evaluate the model on the PASCAL VOC dataset. SMILe introduces two important hyper-parameters,

[3] The default seeds for all our experiments were adapted from `http://dl.yf.io/fs-det/datasets/`

Table 2: Ablation study for the key hyper-parameters in SMILe. The chosen values are underlined and associated performance values are indicated in **bold**.

Parameter	Value	mAP_{base}	mAP_{novel}
Similarity Kernel (S)	Euclidean	84.7	59.4
	Cosine	**88.9**	**61.3**
	RBF	86.1	59.6
η (sim. Kernel = Cosine)	0.0	87.5	59.9
	0.2	88.7	60.2
	0.5	**89.3**	**62.0**
	0.8	86.7	61.1
	1.0	86.1	58.3
λ L_{comb}^{inter} (S = Cosine, η=0.5)	0.5	82.1	57.3
	0.7	86.4	60.1
	1.0	**87.4**	**60.3**
	1.2	87.4	59.9
	1.5	87.1	54.6

similarity kernel S and η. The choice of similarity kernel determines how gradients are calculated in the objective function and they magnitude of S depends on the model parameters θ. We chose the cosine similarity (indicated as Cosine in Tab. 2) metric over others as it achieves the best overall performance. The hyper-parameter η controls the contribution of L_{comb}^{inter} over L_{comb}^{intra} such that their overall contributions add up to 1.0 (100%). We vary the value for η between $\alpha = 0.0$ to $\alpha = 1.0$ and record the variation in performance of the novel classes in Tab. 2. We choose $\eta = 0.5$ for our experiments across all datasets.

Additionally, we introduce the hyper-parameter λ specific to SMILe-GCMI to control the degree of compactness of the feature cluster ensuring sufficient diversity is maintained in the feature space. Experimental results in Tab. 2 indicates that $\lambda \geq 1.0$ is necessary for Graph-Cut in L_{comb}^{intra} to be submodular thus we adopt $\lambda = 1.0$ for our experiments.

4 Proofs for Theorems in SMILe

In this section, we provide the necessary proofs leading to the derivation of the components of L_comb namely, L_{comb}^{inter} and L_{comb}^{intra} for different instantiations of the submodular function $f(A, \theta)$ over any given set A. We restate the theorems as in the main paper for better readability.

4.1 Derivation of SMILe-FLMI

Given $I_f(Q, A) = \sum_{i \in Q} \max_{j \in A} S_{ij}(\theta) + \lambda \sum_{i \in A} \max_{j \in Q} S_{ij}(\theta)$ and $f(A, \theta) = \sum_{i \in \mathcal{T}} \max_{j \in A} S_{ij}(\theta)$ representing the facility-location mutual information function and facility-location submodular function respectively over sets A and Q then, we derive the expressions for SMILe-FLMI as a summation of $L_{comb}^{inter}(\theta)$ and $L_{comb}^{intra}(\theta)$ respectively as depicted in Eq. 6 of the main paper.

Lets first derive the L_{comb}^{intra} from the total information formulation given $f(A, \theta)$ as the underlying submodular function. From the definition of L_{comb}^{intra}, the objective can be derived as $L_{comb}^{intra}(\theta) = \sum_{k \in (C_b \cup C_n)} f(A_k, \theta)$. Substituting the instance of FL $f(A, \theta) = \sum_{i \in \mathcal{V}} \max_{j \in A} S_{ij}(\theta)$ in the equation we get:

$$
\begin{aligned}
L_{comb}^{intra}(\theta) &= \sum_{k=1}^{|C_b \cup C_n|} f(A_k, \theta) \\
&= \sum_{k \in (C_b \cup C_n)} \sum_{i \in \mathcal{T}} \max_{j \in A_k} S_{ij}(\theta) \\
&= \sum_{k \in (C_b \cup C_n)} \sum_{i \in \mathcal{T} \setminus A_k} \max_{j \in A_k} S_{ij}(\theta) + \sum_{k \in (C_b \cup C_n)} \sum_{i \in A_k} \max_{j \in A_k} S_{ij}(\theta) \\
L_{comb}^{intra}(\theta) &= \sum_{k \in (C_b \cup C_n)} \sum_{i \in \mathcal{T} \setminus A_k} \max_{j \in A_k} S_{ij}(\theta) + |\mathcal{T}|
\end{aligned}
$$

Here, $\sum_{i \in A_k} \max_{j \in A_k} S_{ij}(\theta)$ is a constant over the set A_k. Hereafter, we provide the proof for the L_{comb}^{inter} formulation which can be derived from $L_{comb}^{inter}(\theta) = \sum_{\substack{i \in (C_b \cup C_n) \\ j \in C_n : i \neq j}} I_f(A_i, A_j; \theta)$. Given the Submodular Mutual Information function $I_f(Q, A) = \sum_{i \in Q} \max_{j \in A} S_{ij}(\theta) + \lambda \sum_{i \in A} \max_{j \in Q} S_{ij}(\theta)$ over two distinct sets Q and A, we substitute the value of I_f in L_{comb}^{inter}.

$$
\begin{aligned}
L_{comb}^{inter}(\theta) &= \sum_{\substack{k \in (C_b \cup C_n) \\ l \in C_n : k \neq l}} I_f(A_k, A_l; \theta) \\
&= \sum_{k \in (C_b \cup C_n)} \sum_{\substack{l \in C_n \\ k \neq l}} \left[\sum_{i \in A_k} \max_{j \in A_l} S_{ij}(\theta) + \lambda \sum_{i \in A_l} \max_{j \in A_k} S_{ij}(\theta) \right] \\
L_{comb}^{inter}(\theta) &= \sum_{\substack{k \in (C_b \cup C_n) \\ l \in C_n : k \neq l}} \left[\sum_{i \in A_k} \max_{j \in A_l} S_{ij}(\theta) + \lambda \sum_{i \in A_l} \max_{j \in A_k} S_{ij}(\theta) \right]
\end{aligned}
$$

Note that the similarity computed between sets of features depend on the parameters of the model θ. We parallelized the computation of SMILe-FLMI in our implementation using vectorized calculations available in the Pytorch (`https://pytorch.org/`) library.

4.2 Derivation of SMILe-GCMI

From $I_f(Q,A) = 2\lambda \sum_{i \in Q} \sum_{j \in A} S_{ij}(\theta)$ and $f(A,\theta) = \sum_{i \in A} \sum_{j \in \mathcal{T} \setminus A} S_{ij}(\theta) - \lambda \sum_{i,j \in A} S_{ij}(\theta)$ representing the Graph-Cut (GC) mutual information function and Graph-Cut submodular function respectively over sets A and Q then, $L^{inter}_{comb}(\theta)$ and $L^{intra}_{comb}(\theta)$ we derive the expressions for SMILe-GCMI as a summation of $L^{inter}_{comb}(\theta)$ and $L^{intra}_{comb}(\theta)$ respectively as depicted in Eq. 7 in the main paper.

From the definition of $f(A,\theta)$, the SMILe-GCMI (L^{intra}_{comb}) objective can be derived by substituting the instance of GC $f(A_k,\theta)$ in the equation we get:

$$\begin{aligned} L^{intra}_{comb}(\theta) &= \sum_{k \in (C_b \cup C_n)} f(A_k,\theta) \\ &= \sum_{k \in (C_b \cup C_n)} \sum_{i \in A_k} \sum_{j \in \mathcal{T}} S_{ij}(\theta) - \lambda \sum_{i,j \in A_k} S_{ij}(\theta) \\ &= \sum_{k \in (C_b \cup C_n)} \sum_{\substack{i \in A_k \\ j \in \mathcal{T} \setminus A_k}} S_{ij}(\theta) + \sum_{k \in (C_b \cup C_n)} \sum_{\substack{i \in A_k \\ j \in A_k}} S_{ij}(\theta) - \lambda \sum_{i,j \in A_k} S_{ij}(\theta) \end{aligned}$$

Here, the term $\sum_{k \in (C_b \cup C_n)} \sum_{i \in A_k, j \in A_k} S_{ij}(\theta)$ represents a sum of pairwise similarities over all sets in $\mathcal{V}$. Thus, its value is a constant for a fixed training/ evaluation dataset. Using this condition and ignoring the constant term, we can show that:

$$L^{intra}_{comb}(\theta) = \sum_{k \in (C_b \cup C_n)} \sum_{i \in A_k, j \in \mathcal{T} \setminus A_k} S_{ij}(\theta) - \lambda \sum_{i,j \in A_k} S_{ij}(\theta)$$

Hereafter, we provide the proof for the L^{inter}_{comb} formulation which can be derived from $L^{inter}_{comb}(\theta) = \sum_{\substack{i \in (C_b \cup C_n) \\ j \in C_n : i \neq j}} I_f(A_i, A_j; \theta)$. Given the Submodular Mutual Information function $I_f(Q,A) = 2\lambda \sum_{i \in Q} \sum_{j \in A} S_{ij}(\theta)$ over two distinct sets Q and A, we substitute the value of I_f in L^{inter}_{comb}.

$$\begin{aligned} L^{inter}_{comb}(\theta) &= \sum_{\substack{k \in (C_b \cup C_n) \\ l \in C_n : k \neq l}} I_f(A_k, A_l; \theta) \\ &= \sum_{k \in (C_b \cup C_n)} \sum_{\substack{l \in C_n \\ k \neq l}} \left[2\lambda \sum_{i \in A_k} \sum_{j \in A_l} S_{ij}(\theta) \right] \\ L^{inter}_{comb}(\theta) &= \sum_{\substack{k \in (C_b \cup C_n) \\ l \in C_n : k \neq l}} 2\lambda \sum_{\substack{i \in A_k \\ j \in A_l}} S_{ij}(\theta) \end{aligned}$$

From the above formulation, we observe that SMILe-GCMI is computationally inexpensive as compared to SMILe-FLMI, but our experimental results show that SMILe-FLMI outperforms SMILe-GCMI. This is predominantly because the objective function in SMILe-FLMI scales non-linearly with the size of the set $|A_k|$ inherently modelling the imbalance between the already learnt classes C_b and the newly added ones C_n.

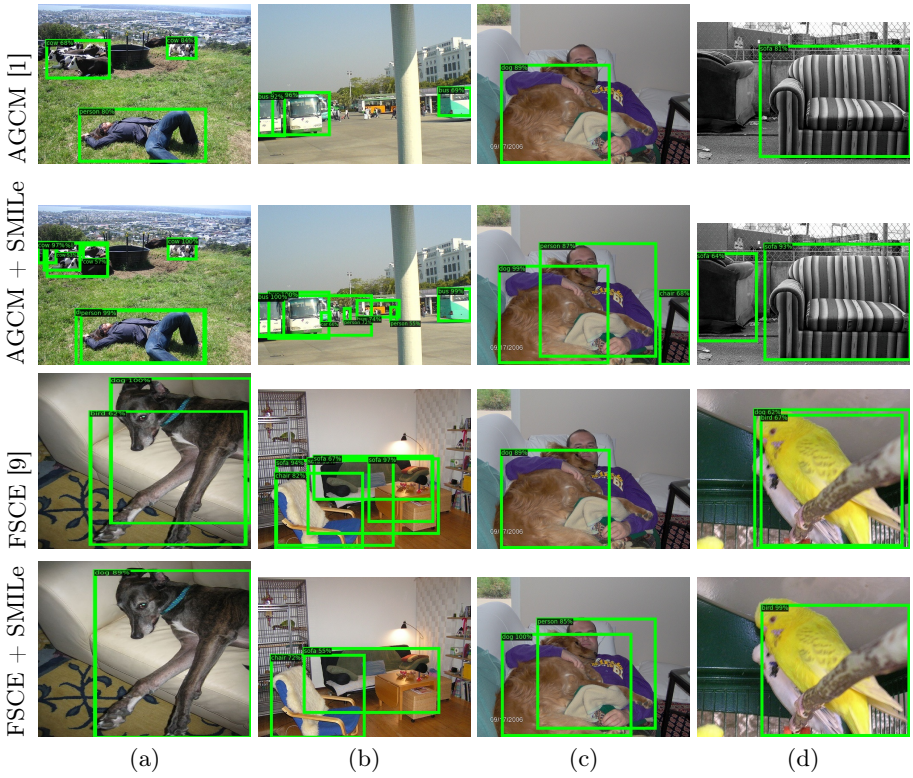

Fig. 1: Qualitative results from SMILe: We contrast the performance of AGCM and FSCE before and after introduction of the Combinatorial formulation introduced in SMILe. We observe significant confusion and forgetting in SoTA approaches FSCE and AGCM while introduction of SMILe overcomes most of these pitfalls.

5 Ablation : Qualitative Results from SMILe Against SoTA

Figure 1 shows qualitative results for our proposed SMILe method on the PASCAL VOC [3]. Due to limited compute resources we conduct experiments on

FSCE and AGCM approaches before and after introduction of the SMILe approach. Figure 1(a) shows that introduction of SMILe is resilient to scale (varying sizes) and occlusion, while Figure 1(c) shows significant base class forgetting in both FSCE and AGCM. Figure 1(b) shows significant catastrophic forgetting in FSCE and AGCM which has also been shown to be overcome by SMILe while Fig. 1(d) demonstrate resilience against color and texture variations. Overall, SMILe handles forgetting and confusion significantly over SoTA approaches while minimizing the degradation in performance of the base classes.

6 Limitations and Future Work

From the experiments proposed in our paper, we demonstrate the generalizability as well as the supremacy of our approach in handling class confusion and forgetting. Although, significant progress has been demonstrated in overcoming confusion and forgetting by SMILe some amount of confusion and forgetting continue to plague this domain. This would definitely be a direction for future research both in FSOD and in combinatorial representation learning. Further, SMILe demonstrates success in the 5/10 shot setting, we observe suboptimal performance in the 1-shot case. This is a plausible direction that the authors would be studying in depth in the near future. In the current setting, novel classes need to be first labelled by human annotators before being served to the SMILe framework. Unfortunately, to rapidly adapt to the open-world setting our model should be able to generalize to unknown Region-of-Interests, which the authors would like to study in future research.

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. 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)
3. 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)
4. 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)
5. 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)
6. 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)
7. 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)
8. 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)
9. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: FSCE: Few-Shot Object Detection Via Contrastive Proposal Encoding. In: CVPR (June 2021)
10. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly Simple Few-Shot Object Detection. In: ICML (2020)