Supplementary Material - Attention Guided Cosine Margin to Overcome Class-Imbalance in Few-Shot Road Object Detection

Ashutosh Agarwal*† Anay Majee ‡ Anbumani Subramanian ‡ Chetan Arora †
IIT Delhi†, Intel Corporation‡
{ashutosh.agarwal,chetan}@cse.iitd.ac.in†, {anay.majee,anbumani.subramanian}@intel.com‡

1. Additional Implementation Details

In this section, we discuss implementation details of existing Few-Shot Object Detection (FSOD) methods for training on India Driving Dataset [4] (IDD). We adapt the open-source implementation of respective methods for most existing approaches and tune their hyper-parameters for IDD data splits. Our proposed AGCM approach is built on top of the Stronger Baseline (SB) proposed in FSCE, which unfreezes the Region Proposal Network (RPN), RoI feature extractor, and Feature Proposal Network (FPN) [2] during the few-shot adaptation stage. However, while adapting this to AGCM for IDD data splits, we do not unfreeze the FPN since few-shot samples tend to cause model overfitting resulting in elevated catastrophic forgetting. For re-implementing the FSCE approach for IDD datasplits we use the contrastive loss weight as 0.5 and the temperature value τ as 0.2 through extensive hyper-parameter tuning on the IDD-OS split. The model has been trained until convergence with a constant learning rate of 0.001 for all split and shot settings.

2. Detailed Architecture

Figure 1 shows a detailed architecture of our proposed method AGCM. RoI proposals are first passed to two fully connected layers (FC1 and FC2) of the RoI Feature Extractor to produce the class-agnostic feature set $P = f(I, \theta_f)$. The feature representations for each proposal $p_i \in P$ are normalized and fused with other similar RoI proposals through the APF module.

The fusion occurs in 3 steps - **(1)** L2 Feature Normalisation (indicated as *Norm* in Figure 1) **(2)** Calculation of attention weights w_{ij} via a cosine similarity metric **(3)** Attentive feature fusion is described in section 3.2.1 of the main paper. The fused proposals are passed to the fully connected linear layer of the classifier $C(\phi, \theta_c)$, which is used for calculation of the Cosine Margin Cross-Entropy Loss $L_{\text{cos-margin}}$ as described in section 3.2.2 of the main paper.

*Work done as an intern at Intel.

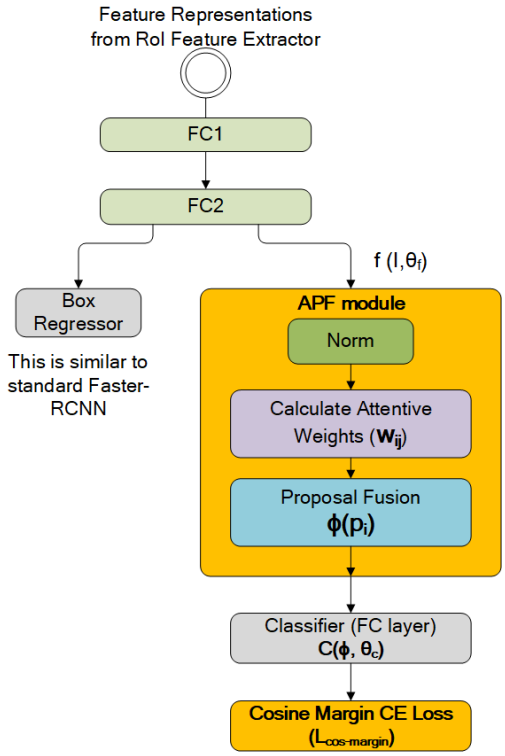

Figure 1: Detailed Architecture of our proposed AGCM method highlighting the key architectural components, a novel APF module, and a cosine margin cross-entropy loss introduced in the classifier head of the Faster-RCNN object detector.

3. Results on PASCAL VOC dataset

This section provides detailed comparative results from the AGCM and FSCE approaches on the PASCAL-VOC dataset [1]. We also discuss the reasoning behind the variation in performance scores obtained from the existing State-of-The-Art (SoTA) FSOD framework, FSCE [3] in this work with respect to the scores reported by the authors. The results from FSCE are a reproduction of the official

Table 1: Seedwise Results on PASCAL VOC for our proposed method AGCM and FSCE. On an average, we outperform FSCE for all shot and split settings.

	K = 10						K = 5						K = 1					
	Split 1		Split 2		Split 3		Split 1		Split 2		Split 3		Split 1		Split 2		Split 3	
Seed	FSCE	AGCM	FSCE	AGCM	FSCE	AGCM	FSCE	AGCM	FSCE	AGCM	FSCE	AGCM	FSCE	AGCM	FSCE	AGCM	FSCE	AGCM
0	57.8	59.9	48.1	49.4	57.7	56.0	57.9	58.5	43.2	42.2	53.2	54.2	40.3	39.3	22.3	23.9	40.1	42.1
9	53.2	57.0	48.8	49.8	46.3	50.0	46.4	53.4	44.4	49.3	48.0	50.8	28.3	28.0	18.9	19.2	10.3	11.6
10	54.2	55.0	43.5	46.3	41.4	45.2	41.0	49.8	40.0	41.4	43.0	42.2	25.9	24.9	9.9	9.5	26.4	25.3
12	54.8	52.4	41.2	44.0	51.9	53.3	51.3	46.2	32.3	34.2	49.8	50.8	33.9	34.5	17.5	16.7	24.9	25.7
13	51.4	53.7	48.2	47.3	44.9	45.9	44.2	47.7	33.7	35.6	46.2	46.1	41.0	40.3	27.3	27.5	32.1	29.1
14	49.1	51.0	40.5	46.3	48.0	50.5	36.8	40.5	31.2	34.7	45.9	46.6	26.2	25.9	6.6	6.3	25.3	27.1
17	58.1	57.5	48.7	50.6	51.8	53.1	42.8	49.7	42.4	45.9	40.6	43.0	21.7	25.2	17.9	19.1	17.5	19.1
18	56.6	55.2	47.0	48.1	52.6	56.2	47.3	47.1	37.3	38.2	46.2	48.0	16.5	15.0	15.0	15.2	18.9	18.2
24	51.2	51.1	36.8	42.7	47.1	48.1	47.4	49.8	27.9	30.9	43.4	45.5	21.7	23.3	9.6	13.4	16.1	18.4
25	54.9	55.4	49.8	45.7	52.4	54.6	46.5	47.2	26.5	32.4	37.6	38.3	26.9	26.6	20.1	21.2	10.6	12.3
Average	54.1	**54.8**	45.3	**47.0**	49.4	**51.5**	46.2	**49.0**	35.9	**38.5**	45.4	**46.5**	28.2	**28.3**	16.5	**17.2**	22.2	**22.9**
Std.dev	±2.96	±**2.87**	±4.45	±**2.53**	±4.72	±**4.31**	±5.74	±**4.72**	±6.49	±**6.07**	±4.49	±**4.69**	±8.00	±**7.71**	±6.37	±**6.41**	±9.43	±**9.06**

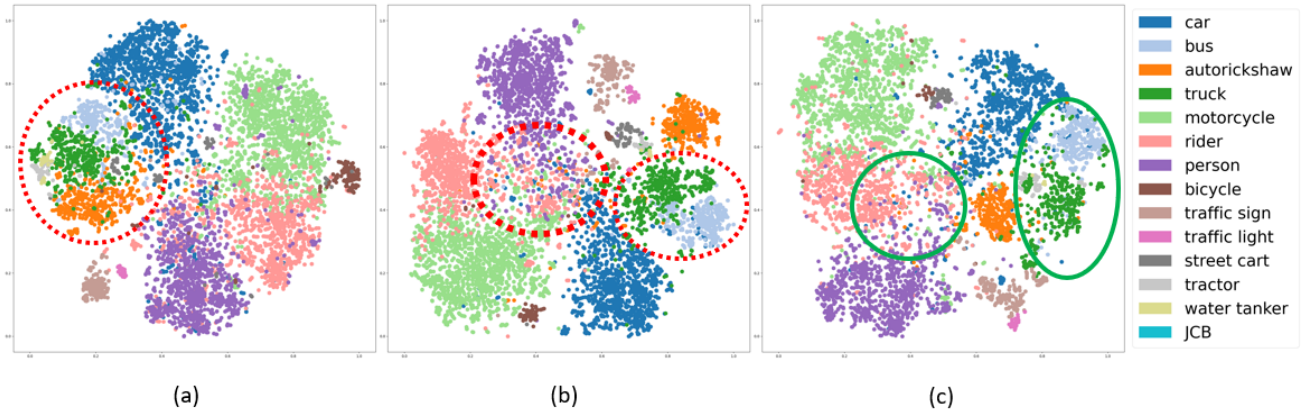

Figure 2: t-SNE plots of feature representations from the classifier head of (a) FsDet [5] (b) FSCE [3] and (c) AGCM (ours) approaches. The plots for FsDet and FSCE show significant overlaps among feature representations (depicted in red) from object classes. Our AGCM technique is able to reduce this overlap by forming tighter and more separated feature clusters (depicted in green).

open-source implementation[1] released by the authors. The implementation details of our proposed AGCM approach are mentioned in detail in section 4.2 of the main paper. The performance of FSOD techniques varies based on the choice of image samples used during the few-shot adaptation stage [5]. This introduces an uncertainty in the performance scores reported by individual methods. Following [5], we evaluate FSCE and AGCM over a series of randomly sampled image sets for 1, 5, and 10-shot settings. We sample these images using randomly selected seeds as shown in Table 1. We also observe a variation in the performance owing to the randomness in the data ingestion pipeline during fine-tuning. For a fair comparison, we fix the seed for data ingestion as a constant value. In the main paper, we reported the best results for FSCE and AGCM and an average across 10 random seeds to alleviate any form of uncertainty associated with the choice of few-shot data samples.

Table 1 shows the results of our proposed algorithm AGCM and our reproduced results for the SoTA metric learner FSCE on the PASCAL-VOC dataset. We see that for a higher shot (K = 5 and K = 10) value, AGCM outperforms FSCE by 1.5 and 2.2 mAP points for 10-shot and 5-shot settings, respectively. However, for the 1-shot setting, we do not see high improvements (0.5 mAP points on average) over our baseline, FSCE. The lack of feature information from a single data sample increases the effect of intra-class variance and inter-class bias in the AGCM approach leading to this drop in novel class performance.

4. Additional Qualitative Results

Figure 3 shows additional qualitative results for our proposed AGCM method on the challenging IDD-OS split. We show that our AGCM approach can detect objects in low-

[1] https://github.com/MegviiDetection/FSCE

light conditions and is resistant to occlusions. It can be seen that our method shows relatively lower catastrophic forgetting by predicting instances of the base classes *person*, *motorcycle* etc., alongside novel classes like street cart, water tanker, etc., after few-shot adaptation with significant confidence. We also quantitatively prove this finding in section 5.4 in the main paper.

We also demonstrate few example predictions from the IDD dataset (depicted in the bottom row of figure 3) where our proposed AGCM approach fails to overcome the challenges of class confusion and catastrophic forgetting. As mentioned in the main paper, this can be attributed to a significant feature representation sharing among object classes.

5. Comparison of Class confusion

Figure 2 shows a t-SNE visualisation of feature representations for FsDet, FSCE and AGCM. We observe a large overlap between the feature representation of object classes in FsDet. Although contrastive learning in FSCE can produce better feature clusters, it does not create sufficient inter-class margin among classes causing overlaps between confusing classes such as bus, autorickshaw, and truck. Our proposed cosine margin Cross-Entropy loss encourages inter-class separation and therefore reduces class confusion among such classes.

References

[1] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object classes (VOC) Challenge. *IJCV*, pages 303–338, 2010.

[2] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. In *CVPR*.

[3] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *CVPR*, June 2021.

[4] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. V. Jawahar. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *WACV*, pages 1743–1751, 2019.

[5] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *ICML*, 2020.