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‡

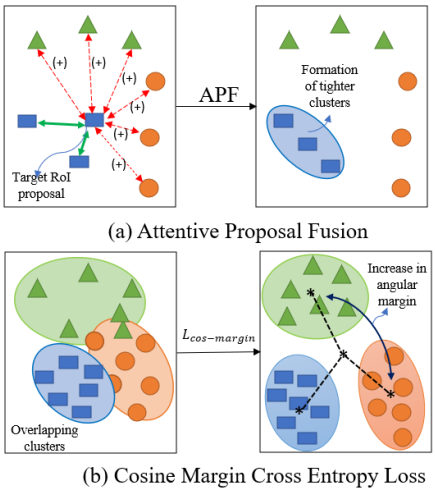
Abstract

*Few-Shot Object Detectors (FSOD) are tasked to localize and classify objects in an image given only a few data samples. Recent trends in FSOD research show the adoption of metric and meta-learning techniques, which are prone to catastrophic forgetting and class confusion. To overcome these pitfalls in metric learning based FSOD techniques, we introduce an **A**ttention **G**uided **C**osine **M**argin (**AGCM**) that facilitates the creation of tighter and well separated class-specific feature clusters in the classification head of the object detector. The Attentive Proposal Fusion (APF) module introduced in AGCM minimizes catastrophic forgetting by reducing the intra-class variance among co-occurring classes. At the same time, the Cosine Margin penalty in AGCM increases the angular margin between confusing classes to overcome the challenge of class confusion between already learned (base) and newly added (novel) classes. We conduct our experiments on the India Driving Dataset (IDD), which presents a real-world class-imbalanced setting alongside popular FSOD benchmark PASCAL-VOC. Our method outperforms existing approaches by up to 6.4 mAP points on the IDD-OS and up to 2.0 mAP points on the IDD-10 splits for the 10-shot setting. On the PASCAL-VOC dataset, we outperform existing approaches by up to 4.9 mAP points.*

1. Introduction

Deep Convolution Neural networks (ConvNets) trained on large-scale image datasets [4, 17], have shown exemplary performance on tasks like classification and object detection [10, 24, 25]. A noticeable pitfall in ConvNets is the requirement of large-scale annotated datasets to achieve State-of-The-Art (SoTA) performance which is both expensive and labor-intensive to acquire.

Recent developments in Machine Learning research have shown significant progress in few-shot learning, especially for image recognition [8, 9, 22, 26, 27, 29, 31] tasks where algorithms learn to recognize images from limited (few-shot) data samples. On the contrary, Few-Shot Object Detection (FSOD) emerges as a relatively unexplored and complex field as it encompasses both localization and recognition tasks.



(a) Attentive Proposal Fusion

(b) Cosine Margin Cross Entropy Loss

Figure 1: Overview of our proposed Attention Guided Cosine Margin approach (a) Visually similar classes are driven closer using our Attentive Proposal Fusion Module (APF). (+) represents that the distance between corresponding RoI proposals should be increased. (b) Our cosine margin cross-entropy loss increases the inter-class margin among classes.

Early attempts in FSOD have been made by drawing inspiration from two primary learning strategies in image classification - Meta-Learning [12, 15, 36, 34] and Metric Learning [28, 32, 37]. Benchmark experiments conducted by these works show that metric learners significantly outperform meta-learners [33] in adapting to few-shot data. However, the success of metric learners is seldom overshadowed by two dominant issues - *class confusion* and *Catastrophic forgetting*. Class confusion refers to misclassifying a predicted Region of Interest (RoI) as an incorrect class

*Work done as an intern at Intel Corporation

label. This confusion is commonly observed among objects belonging to the newly added (novel) classes, which are classified as one or more already learned (base) classes. Catastrophic forgetting refers to the degradation in performance of the base classes while adapting to novel classes. The issues mentioned above become more evident in real-world scenarios such as autonomous driving [2, 30] where only a few data samples are available for detecting less-occurring road objects with significant variations in structure and orientations.

Through extensive experimentation on several SoTA FSOD methods, we observe significant overlaps between feature representations of the base and novel classes. This overlap can be attributed to increasing class confusion in FSOD. On the other hand, catastrophic forgetting among base classes is a result of FSOD techniques [28, 32] overfitting to few-shot data samples.

We propose a metric-learning based **A**ttention **G**uided **C**osine **M**argin (**AGCM**) approach that exploits the overlapping features among RoI proposals in FSOD to create compact and well-separated feature clusters. As shown in Figure 1, our novel Attentive Proposal Fusion (APF) module computes the similarity in features between RoI proposals and assigns higher attentive weights to similar RoIs without referring to the class labels. Since similar RoIs have a high likelihood of belonging to the same class, such feature representations are driven closer in the embedding space, thus forming tighter clusters. APF also ensures that the object detector assigns equal representation to base and novel classes, resulting in reduced catastrophic forgetting. We also introduce a cosine margin cross-entropy loss ($L_{\text{cos-margin}}$ in Figure 1) that overcomes the impact of class confusion by increasing the angular margin between object classes.

Existing works on FSOD demonstrate their performance on canonical benchmarks like PASCAL-VOC [5], and MS-COCO [17] which do not represent the real-world scenarios leading to poor performance during deployment in challenging domains such as autonomous driving. On the contrary, we demonstrate the performance of our approach on the recently introduced benchmark in FSOD, few-shot India Driving Dataset [20] as it presents a real-world, class-imbalanced setting with large intra-class variance and inter-class bias [30]. The main contributions of our work can be summarized as:

- We introduce a simple and lightweight metric learning based FSOD technique, Attention Guided Cosine Margin (AGCM), to overcome class confusion and catastrophic forgetting in driving scenes.
- We introduce a parameterless Attentive Proposal Fusion module (APF) and a Cosine Margin Cross-Entropy loss in AGCM to retain feature information from base classes while generalizing to novel classes.
- We demonstrate upto 10% reduction in class confusion and 18% improvement in catastrophic forgetting while achieving SoTA performance on the challenging India Driving Dataset (IDD) [30] and other FSOD benchmarks like PASCAL-VOC [5].

2. Related Work

2.1. Few-Shot Object Detection

Classical approaches in FSOD adopt a traditional fine-tuning strategy [1], or a distance metric learner [13] to extend the features of the already learned (base) classes to the newly added (novel) classes. Recent approaches in FSOD adopt meta-learning techniques on standard object detection methods using episodic training [12, 34, 36] to learn class-specific feature sets to discriminate among classes. Meta-Reweight [12] and MetaRCNN [36] learns an additional feature extractor network that converts class agnostic features from the RoI head of the object detector to class-specific features. Add-Info [34] introduces the feature difference between meta-train (support) and meta-test (query) images as additional features, while [6] learns an Attention-based Region Proposal Network (RPN) to guide a relation network [29] to learn discriminative features for each class. Very recent approaches in meta-learning like [38] encourages sharing of information between support and query images to enhance class-specific feature sets, while CME [15] establishes an equilibrium between class margins to reduce class confusion and demonstrates better generalization to novel classes. A characteristic feature of meta learners is the use of attention mechanisms [6, 39] to identify the most discriminative features for each class. This allows meta learners to retain the knowledge of base classes while generalizing to novel classes.

Despite their success in retaining the knowledge of base classes, meta-learning approaches are compute and memory-intensive. Surprisingly, metric learning strategies provide a better generalization to novel objects without any additional overheads. FsDet [32] learns generalizable feature embeddings by introducing a cosine-similarity-based classifier. FSCE [28] adopts a contrastive training strategy while SRR-FSD [40] uses the semantic relationships between word embeddings from category labels to show improvements on novel class performance. PNPDet [37] decouples the base and novel class predictors and learns a cosine-similarity classifier that partially resolves catastrophic forgetting and class confusion. Unfortunately, metric learners suffer from extreme catastrophic forgetting as they tend to overfit on the novel classes. GFSD [7] proposes a Bias-Balanced RPN to prevent overfitting on metric learners and introduces a Re-detector network that decouples the base and novel class predictors. Although this

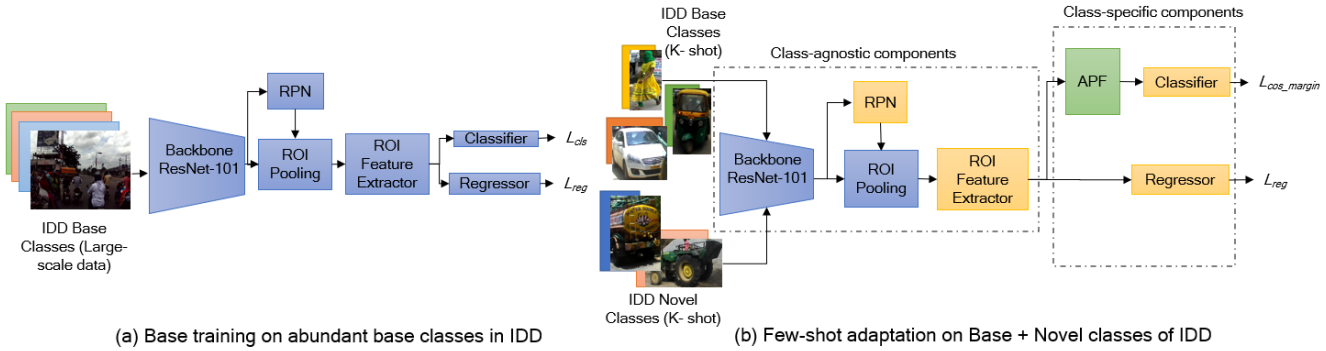


Figure 2: **The architecture of our proposed AGCM approach:** AGCM follows a metric learning strategy and is applied during the few-shot adaptation stage (b) after base training (a), on K shot samples from base and novel categories. We introduce an Attentive Proposal Fusion (APF) module and a cosine margin cross-entropy loss to overcome class confusion and catastrophic forgetting in FSOD.

technique reduces the impact of catastrophic forgetting, it fails to generalize to novel classes. Our work introduces attention-based proposal level fusion into metric learning based FSOD technique to help retain information from base classes, preventing catastrophic forgetting.

The authors in [20] demonstrate the application of FSOD in the context of autonomous driving to detect less-occurring road objects. Our work adopts this problem definition and demonstrates a reduction in catastrophic forgetting while showing significant improvements in the performance of novel classes.

2.2. Margin based Feature learning

Margin-based learning has been applied to various computer vision tasks [19, 21, 35] to better discriminate between objects that show a significant overlap in visual features. Such penalties have proven to be effective in reducing class confusion for the few-shot classification [14, 18] task by introducing an additional angular margin between feature clusters. While similar approaches have been recently adopted in meta-learning based FSOD techniques [15, 38], the margin-based penalty is yet to be explored for metric learners. To the best of our knowledge, we are the first to introduce a simple and effective margin-based penalty in metric learning based FSOD techniques through the Cosine Margin Cross-Entropy loss described in section 3.2.2.

3. Method

In this section, we define the problem for Few-Shot Object Detection and describe the architecture of our proposed Attention Guided Cosine Margin (AGCM) approach.

3.1. Problem Definition

We define a proposal based few-shot object detector $h(I, \theta)$ consisting a class-agnostic component $f(I, \theta_f)$ and a class-specific component $C(f, \theta_c)$ as shown in figure 2(b), such that $h(I, \theta) = C(f(I, \theta_f), \theta_c)$. Here, I represents the input images and θ, θ_f and θ_c represents the respective model parameters for the components of the few-shot object detector. We define a metric learning based FSOD training strategy as in [20] which proceeds in two stages: *base training* and *few-shot adaptation*. During base training $h(I, \theta)$ learns to detect objects from base classes (C_{base}) using a large scale dataset D_{base}. In the few-shot adaptation stage, $h(I, \theta)$ is fine-tuned using images in D_{novel} consisting of classes $C_{base} \cup C_{novel}$ with only K instances from N classes, such that $|C_{base} \cup C_{novel}| = N$. The goal for $h(I, \theta)$ is to boost performance on novel classes in D_{novel} with minimal degradation in performance of classes in D_{base}.

3.2. AGCM: Attention Guided Cosine Margin

The proposed Attention Guided Cosine Margin (AGCM) approach adopts a novel metric learning strategy to reduce the intra-class variance and inter-class bias among object classes by encouraging orthogonality among class-specific feature clusters [23]. As shown in Figure 2(b) the AGCM is applied only during the few-shot adaptation stage to the output of the class-agnostic branch of the object detector $f(I, \theta_f)$ to guide the class-specific component $C(f, \theta_c)$ through two key components. We first apply a novel Attentive Proposal Fusion (APF) to the feature representations of individual RoI proposals in $P = f(I, \theta_f)$. The feature information in each RoI proposal P_i is propagated across all proposals $P_j \in P$ and is fused with those that have high visual similarity with P_i in a label-free fashion. Secondly, we introduce a Cosine Margin Cross-Entropy loss term to the classification head of the object detector $C(f, \theta_c)$. This loss term maximizes the angular separation between feature clusters to reduce inter-class bias among classes. We

describe the formulations of the APF and Cosine Margin Cross-Entropy loss in sections 3.2.1 and 3.2.2 respectively. The combined effect of these two modules results in a significant reduction in class confusion and catastrophic forgetting, as shown by our experiments in section 4.

3.2.1 APF: Attentive Proposal Fusion Module

The scarcity of data samples in FSOD techniques leads to the formation of non-discriminative feature sets, especially for the novel classes. The bias associated with few-shot data has been identified in [3] as *sample bias* for image recognition tasks. The authors in [3] propose a transductive meta-learning strategy by propagating feature information between labeled few-shot samples (support set) and unlabelled test data samples (query set). We adopt a similar direction through the Attentive Proposal Fusion (APF) with modifications towards metric learning based FSOD techniques.

The class-agnostic component of the object detector $f(I, \theta_f)$ produces feature representations from M RoI proposals, denoted by P. Our proposed APF module is applied to individual RoI proposals in P to maximize the class-specific feature information by fusing the low-level features from RoI proposals $p_i \in P$ with the weighted sum of the remaining $M-1$ proposals as described in equation 1. Here, $\Phi(p_i)$ represents the proposal p_i after feature fusion and w_{ij} is the attentive weight between the i^{th} and the j^{th} RoI proposal.

$$\Phi(p_i) = \alpha \cdot p_i + (1-\alpha) \sum_{j \in \mathbb{P}, j \neq i} w_{ij} \cdot p_j \tag{1}$$

As described in (2) the attentive weights (w_{ij}) represents the likelihood of the features in the i_{th} RoI proposal to be similar to the features in the j_{th} proposal. It involves a non-linear similarity (cosine similarity [9] in our case) between p_i and p_j denoted by $cos(p_i, p_j)$. The choice of this metric is described in detail in section 5.2.

$$w_{ij} = \frac{e^{cos(p_i, pj)}}{\sum_{k \neq i, k \in \mathbb{P}} e^{cos(p_i, pk)}} \tag{2}$$

The formulation of $\Phi(p_i)$ introduces a hyper-parameter α which controls the proportion of low-level features that are fused into p_i from remaining RoI proposals. The value of α is always kept in the range $[0.5, 1.0]$ to encourage the retention of a significant portion of the features of the original RoI. More details on the choice of α is provided in section 5.2.

The information exchange among RoI proposals encourages the grouping of similar feature representations without the ground truth label information. This facilitates the reduction in intra-class bias and chances of model overfitting as all classes in the training dataset are equally represented in the embedding space. Consequently, we observe a diminishing effect on catastrophic forgetting of base classes as shown in section 5.4.

3.2.2 Cosine Margin Cross-Entropy Loss

Although the application of label-free feature fusion (APF module) helps in forming tighter feature clusters, it may result in the clustering of features from heterogeneous classes that show high visual similarities. It also fails to ensure sufficient margin among co-occurring object classes like *motorcycle* and *rider*, leading to elevated class confusion.

Based on the recent success of margin based penalties in auxiliary vision tasks (section 2.2), we introduce a negative angular margin based loss function in AGCM with suitable modifications for metric learning based FSOD techniques. In contrast to a positive margin, a negative margin helps establish an equilibrium between the distinguishability of classes and the performance of novel classes [18].

We apply the cosine margin-based objective in the few-shot adaptation stage, to the output logits of the classification head in the FSOD model, $Z = C(\phi, \theta_c)$, where θ_c represents the parameters of the classifier head and ϕ is obtained by applying APF module on the RoI features from the class-agnostic branch $f(I, \theta_f)$. The objective function described in equation 3 through 4 maximizes the log-likelihood of the angular distance between the logit $z_i \in Z$ corresponding to the ground truth label y_i and the normalized weight vector of the corresponding class W_{y_i}.

$$L_{\text{cos-margin}} = -\frac{1}{M} \sum_{i=1}^{M} l(z_i) \tag{3}$$

$$l_{z_i} = \log \frac{e^{\beta\left(\cos\left(z_i, W_{y_i}\right) - \mathbb{1}_{y_i \neq back} m\right)}}{e^{\beta\left(\cos\left(z_i, W_{y_i}\right) - \mathbb{1}_{y_i \neq back} m\right)} + \sum_{j=1, j \neq y_i}^{N} e^{\beta \cos(z_i, W_j)}} \tag{4}$$

An angular margin m is applied to this objective to increase the separation between feature clusters, and a scaling factor β is introduced which is set to a constant value of 20 [18]. Also, we do not apply the angular margin to logits of the *background* class to prevent loss of information during model training as it might contain features belonging to one or more object classes in $C_{base} \cup C_{novel}$. The choice of the value of margin m is described in section 5.2.

3.2.3 Training Procedure

As defined in section 3.1 the model $h(I, \theta)$, involves a Faster-RCNN [25] based object detector and trained in two distinct stages. We adopt the training strategy of FsDet [32] during the base training stage, and train $h(I, \theta)$ till convergence. We use the standard loss function used in [25]

Table 1: **Results on Few-Shot India Driving Dataset:** Few-shot object detection performance (mAP_{50}) on IDD-OS and IDD-10 splits from India Driving Dataset using 5 and 10 shot samples.

Data-split	**IDD-OS**				**IDD-10 (Split 1)**				**IDD-10 (Split 2)**			
Shots (K)	*K=5*		*K=10*		*K=5*		*K=10*		*K=5*		*K=10*	
Method \ **Metric**	mAP_{base}	mAP_{novel}	mAP_{base}	mAP_{novel}	mAP_{base}	mAP_{novel}	mAP_{base}	mAP_{novel}	mAP_{base}	mAP_{novel}	mAP_{base}	mAP_{novel}
Meta-RCNN [36]	24.1	4.3	24.0	6.4	23.2	5.7	24.6	7.8	18.1	7.4	18.2	6.7
Add-Info [34]	36.4	18.2	37.1	28.8	33.5	5.2	33.7	10.0	31.3	7.7	32.1	9.5
FsDet w/ cos [32]	38.2	23.6	47.8	39.8	33.5	13.1	31.2	22.1	34.2	14.8	39.7	22.8
FSCE [28]	38.1	39.1	45.5	51.6	23.6	9.2	31.3	16.4	30.6	9.1	37.7	14.7
AGCM (ours)	**42.1**	**45.5**	**51.5**	**58.0**	**37.2**	**16.0**	**45.0**	**22.1**	**36.2**	**15.2**	**42.3**	**24.8**

(a) Street Cart (b) Excavator (c) Water Tanker (d) Tractor

Figure 3: **Qualitative results from the few-shot India Driving Dataset:** We contrast the performance of AGCM against SoTA FSOD approach, FSCE for novel classes in the IDD-OS split for the 10-shot setting. FSCE suffers from extreme catastrophic forgetting and is unable to adapt to large intra-class and inter-class variations in IDD. Such issues are shown to have been overcome by the proposed AGCM approach.

comprising of a binary cross-entropy loss at the Region Proposal Network (RPN) to separate foreground and background proposals L_{rpn}, a cross-entropy loss for bounding box classifier L_{cls} and a smoothed L1 loss to localize the bounding box deltas L_{reg}.

In the few-shot adaptation stage, we adopt the stronger baseline presented by FSCE [28] in which the network backbone remains frozen, the number of proposals generated by the RPN is doubled, and the number of RoI features used for loss computation is halved. This is done to facilitate the incorporation of the low-confidence predictions from the novel classes during the initial training iterations. We add our APF module to the classifier head of $h(I, \theta)$ and replace the cross-entropy loss L_{cls} with our proposed cosine margin cross-entropy loss $L_{\text{cos-margin}}$ as shown in (5).

$$L = L_{rpn} + L_{\text{cos-margin}} + L_{reg} \tag{5}$$

4. Experiments

In this section, we describe our experimental setup and compare the results of our proposed method with existing FSOD techniques on multiple benchmark datasets. We adopt the standard evaluation criterion in FSOD [32, 12] and report the Mean Average Precision (mAP) at 50% Intersection Over Union (IoU) for all our experiments.

4.1. Datasets

We evaluate our proposed AGCM approach on two few-shot object detection datasets - India Driving Dataset (IDD) [30] and PASCAL-VOC [5] datasets.

Indian Driving Dataset (IDD) comprises of 15 object classes in the IDD-Detection dataset consisting of driving scenes on Indian roads. We adopt the data splits proposed in [20] to evaluate our AGCM approach. The dataset consists of two few-shot data splits -

- **IDD-OS** consists of 14 classes representing an open-world deployment setting with 10 base classes and 4 novel classes. The novel classes (*Tractor*, *Street Cart*, *Water tanker* and *Excavator (JCB)*) have been obtained by expanding on the *vehicle fallback* category in IDD.

- **IDD-10** consists of 10 classes forming 2 few-shot data splits. Each split consists of 7 base classes and 3 randomly chosen novel classes. The authors of [20] create two representative splits, referred to as split 1 (*bicycle*, *bus* and *truck* / others) and split 2 (*auto-rickshaw*, *motorcycle*, *truck*/ others) based on the choice of novel

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 different evaluation strategy (N-way, K-shot meta testing).

Method	Meta/ Metric Learner	Backbone	**Novel Split 1**			**Novel Split 2**			**Novel Split 3**		
			K=1	5	10	1	5	10	1	5	10
† Meta-RCNN [36]	Meta	FRCN-R101	19.9	45.7	51.5	10.4	34.8	45.4	14.3	41.2	48.1
†Meta-Reweight [12]	Meta	YOLO V2	14.8	33.9	47.2	15.7	30.1	40.5	21.3	42.8	45.9
†MetaDet [33]	Meta	FRCN-R101	18.9	36.8	49.6	21.8	31.7	43.0	20.6	43.9	44.1
†Add-Info [34]	Meta	FRCN-R101	24.2	49.1	57.4	21.6	37.0	45.7	21.2	43.8	49.6
†CME [15]	Meta	YOLO V2	17.8	44.8	47.5	12.7	33.7	40.0	15.7	44.9	48.8
PNPDet [37]	Metric	DLA-34	18.2	-	41.0	16.6	-	36.4	18.9	-	36.2
FsDet w/ FC [32]	Metric	FRCN-R101	36.8	55.7	57.0	18.2	35.5	39.0	27.7	48.7	50.2
FsDet w/ cos [32]	Metric	FRCN-R101	39.8	55.7	56.0	23.5	35.1	39.1	30.8	49.5	49.8
FSCE [28]	Metric	FRCN-R101	**41.0**	57.9	57.8	27.3	44.4	49.8	40.1	53.2	57.7
AGCM (ours)	Metric	FRCN-R101	40.3	**58.5**	**59.9**	**27.5**	**49.3**	**50.6**	**42.1**	**54.2**	**58.2**
∗FsDet w/ cos [32]	Metric	FRCN-R101	25.3	47.9	52.8	**18.3**	34.1	39.5	17.9	40.8	45.6
∗FSCE [28]	Metric	FRCN-R101	28.2	46.2	54.1	16.5	35.9	45.3	22.2	45.4	49.4
∗**AGCM (ours)**	Metric	FRCN-R101	**28.3**	**49.0**	**54.8**	17.2	**38.5**	**47.0**	**22.9**	**46.5**	**51.5**

classes. We adopt these splits in our work.

We evaluate our approach on the complete validation set of IDD for 5 and 10 shot settings.

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 [12], 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.

4.2. Experimental Setup

The architecture of the proposed AGCM is based on the Faster-RCNN [25] model with a ResNet-101 [11] and Feature Pyramidal Network [16] based backbone. For IDD, the input batch size to the network is set to 2 and 6 in the base training and few-shot adaptation stages respectively. However, for PASCAL-VOC, a batch size of 16 is used for both stages. The input resolution is set to 1920 x 1080 pixels for data splits in IDD, while it is set to 800 x 600 pixels for PASCAL-VOC. Following the training procedure described in section 3.2.3, we train our model till convergence with a learning rate of 0.001 for both base and few-shot adaptation stages. For IDD, base-training is done for 50k iterations with a pretrained imagenet [4] backbone, while the training procedure of FSCE [28] is followed for PASCAL-VOC. Standard data augmentation like horizontal flip and random crop are applied for both datasets. During the few-shot adaptation stage, we adopt the stronger baseline of FSCE and set the number of RoI proposals to 2000 and the number of RPN proposals to 256. The hyper-parameters used in the formulation of AGCM, namely α, margin (m), and distance, are chosen through ablation experiments described in section 5. Results from existing methods are a reproduction of the algorithm from publicly available codebases along with hyper-parameter tuning on IDD datasplits. Unlike other FSOD benchmarks, all our experiments are performed on a single GPU with 12GB memory. More details can be found in the supplementary material.

4.3. Results on India Driving Dataset

We follow the benchmark experiments in [20] and compare the performance of our AGCM approach against State-of-The-Art (SoTA) meta [34, 36], and metric learners [32] on IDD-OS and IDD-10 splits. Additionally, we extend this benchmark by reimplementing the results of the current SoTA approach in FSOD, FSCE [28] on IDD datasplits. Table 1 records both the base and novel class performance of various approaches in contrast to our AGCM appraoch. For IDD-10 splits, our AGCM outperforms existing SoTA methods by an average of 1.5 mAP points in split-1 and 1.2 mAP points in split-2 on novel classes. For, IDD-OS split, AGCM outperforms the SoTA metric learner, FSCE, by 6.4 and 6 mAP points for the 5 and 10 shot settings, respectively. Alongside the significant improvements in novel class performance, our AGCM approach achieves the highest retention in base class performance, which effectively overcomes catastrophic forgetting. This is further described in section 5.4. Although FSCE has proven to be effective against class confusion and catastrophic forgetting for canonical datasets, there exists a large performance gap between FSCE and AGCM on the IDD datasplits. This can be attributed to the contrastive training strategy in FSCE resulting in elimination of discriminative features for con-

Table 3: Ablation on various components of the proposed AGCM approach.

Method	Stronger Baseline [28]	APF (Sec. 3.2.1)	Cosine Margin CE loss	mAP_{novel} 5-shot	mAP_{novel} 10-shot
FsDet w/ cos	-	-	-	23.6	39.8
FSCE	✓			38.7	51.3
AGCM (ours)	✓	✓		43.3	54.9
	✓	✓	✓	**45.5**	**58.0**

Table 4: Ablation for the effect of key hyper-parameters (α, distance and m) on novel class performance in IDD-OS. The chosen values for the AGCM approach is underlined and associated performance values are indicated in **bold**.

Parameter	Value	mAP_{base}	mAP_{novel}
α (Distance = Euclidean)	0.5	48.9	44.8
	0.7	52.2	52.9
	0.8	**52.7**	**53.9**
	0.9	52.2	54.4
	1.0	50.5	52.7
Distance ($\alpha = 0.8$)	Euclidean	52.7	53.9
	Cosine	**52.7**	**54.9**
	Pearson	52.1	54.8
m (Distance = Cosine, α = 0.8)	0.0	52.7	54.9
	0.1	52.0	56.1
	0.2	**51.5**	**57.9**
	0.4	50.9	53.8
	0.8	49.1	40.1
	1.0	48.5	43.3

fusing road objects.

Figure 3 demonstrates a qualitative analysis of our approach against the SoTA metric learner FSCE on the IDD-OS split in the 10-shot setting. As observed from the figure, the FSCE approach suffers from significant catastrophic forgetting (as shown in figure 3(a)) and is unable to detect incomplete (Water tanker in figure 3(d)) or obscure objects (Excavator in figure 3(b)). Unlike FSCE, the AGCM approach can retain most base class predictions and is invariant to large intra-class variances in IDD.

4.4. Results on PASCAL-VOC dataset

Table 2 records the results obtained from our AGCM approach on novel splits of the PASCAL-VOC dataset and contrasts it against SoTA FSOD techniques. Our method outperforms SoTA approaches on almost all few-shot settings with a maximum improvement of 4.9 mAP points in split-2 for the 5-shot setting. However, we do not achieve high gains for very low shot settings (1-shot) as the model suffers significant inter-class bias and intra-class variance. More experimental details are provided in the supplementary material.

5. Ablation

In this section, we conduct ablation experiments on the challenging IDD-OS split to qualify the contributions of various components and hyper-parameters in our proposed AGCM approach.

Table 5: Ablation experiment on catastrophic forgetting of base classes on IDD-OS split in the 10-shot setting. The mAP_{base} before few-shot adaption is 63.4 mAP points.

Method	mAP_{base}	mAP_{novel}	% drop (↓)
FRCNN-ft (only base classes)	63.4	-	-
FsDet w/ cos	47.8	39.8	24.6
FSCE	45.5	51.6	28.2
AGCM (ours)	**51.5**	**58.0**	**18.8**

5.1. Components of the AGCM Architecture

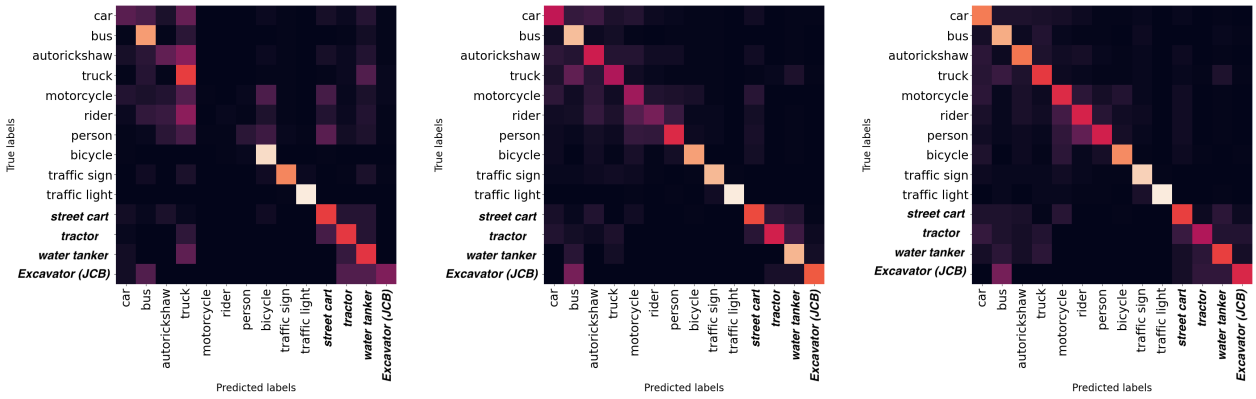
The AGCM approach consists of three main components. First, we adopt the stronger baseline of FSCE [28] which facilitates the inclusion of low-confidence proposals of the novel classes resulting in a significant performance gain over the best performing architecture demonstrated in [20] on IDD-OS. Secondly, our proposed APF module (refer section 3.2.1) reduces the effect of catastrophic forgetting by encouraging the formation of tighter class-specific feature clusters through attentive re-weighting of the RoI proposals. Finally, the cosine margin cross-entropy loss reduces the inter-class bias by increasing the angular margin between feature clusters. It ensures a reduction in confusion among object classes that share a large portion of low-level features. The quantitative contributions of each component is tabulated in Table 3.

5.2. Ablation on key hyper-parameters in AGCM

We perform ablation on various hyper-parameters introduced in our approach and derive their values which lead to the best possible novel class performance during the few-shot adaptation stage.

Hyper-parameters of the APF Module : The hyper-parameter α introduced in section 3.2.1 controls the ratio of the contribution of the chosen RoI with respect to other RoIs in the APF module. We vary the value for α between $\alpha = 0.5$ to $\alpha = 1.0$ and record the variation in performance of the novel classes in Table 4. For smaller values of α, there is a loss of distinctiveness for a feature proposal, and therefore, we see a loss in performance for both base and novel classes. On the other hand, for higher values of α, no information propagation happens among the RoI proposals, which increases class confusion and deteriorates the performance on the base class. We thus chose $\alpha = 0.8$ for our experiments across all datasets.

Attentive weights (w_{ij}) computed for each RoI proposal through equation 2 are calculated through a learnable metric. We ablate this metric in table 4 and record the variations in base and novel class performance while maintaining α at



(a) FsDet (Confusion = 68 %) (a) FSCE (Confusion = 56 %) (a) AGCM (Confusion = 46 %)

Figure 4: Confusion Matrix plot for the proposed AGCM technique. Our method shows a significant reduction (10%) in class confusion between base and novel classes as compared to SoTA metric learning based FSOD techniques FSCE and FsDet. Class names mentioned in ***bold and italics*** represent the novel classes in IDD-OS.

0.8. We chose the cosine similarity metric over others as it achieves the best overall performance.

Hyper-parameters of the Cosine Margin Cross-Entropy Loss : As shown in Table 4, we vary the value of m in the range of [0,1] and observe an increase in novel class performance between $m \geq 0$ to $m \leq 0.2$ followed by a decrease in both base and novel class performance from $m > 0.2$ to $m \leq 1.0$. Consequently, we adopt the value of m as 0.2, which results in the highest overall performance gains for all our experiments across datasets. Although we observe a large gain in novel class performance (4 mAP points), a small drop (1 mAP point) is observed in the base class performance as the angular margin increases the inter-class bias among classes.

5.3. Class Confusion Among Road Objects

Figure 4 shows the confusion matrix for our proposed AGCM in contrast with FSCE and FsDet approaches on all classes in IDD-OS for the 10-shot setting. FsDet shows a large confusion of 68.4% while FSCE has 56% confusion. AGCM achieves the least confusion of 46%. Although FsDet can discriminate between the novel classes, it shows significant confusion among base classes with large intra-class variance like *car*, *bus* and *truck*. The contrastive training strategy adopted by FSCE can reduce confusion between such classes but fails to overcome the inter-class bias between co-occurring classes like *motorcycle*, *person*, and *rider*. AGCM overcomes the intra-class variance through proposal fusion (APF) while encouraging inter-class separation through margin penalties, reducing class confusion among classes. We also visualize a t-SNE plot of feature representations in the supplementary material that shows how AGCM helps in reducing overlap between class-specific clusters.

5.4. Catastrophic Forgetting of Base Classes

This section quantifies the drop in base class performance for multiple metric learning techniques and shows that our proposed AGCM approach achieves minimum degradation in base class performance while boosting the performance of novel classes. Results from a Faster-RCNN model with a ResNet-101 backbone trained on 10 base classes in IDD-OS (referred as FRCNN-ft in table 5) is used as the roofline for all evaluations. The results of this experiment after the few-shot adaptation stage (in 10-shot setting) are demonstrated through table 5. Our method achieves the least degradation in base class performance of 18.8% while obtaining the highest base and novel class performance.

6. Conclusion

In this work, we introduced a novel FSOD technique, Attention Guided Cosine Margin (AGCM), to overcome the class imbalance in Few-Shot Road Object Detection. Our method achieves State-of-The-Art (SoTA) results on all the splits of India Driving Dataset, outperforming the SoTA metric learners by up to 6.4 mAP points in IDD-OS split 10-shot setting. AGCM also generalizes to standard FSOD benchmarks like PASCAL-VOC, where we outperform SoTA approaches by up to 4.9 mAP points. Our proposed Attention Proposal Fusion (APF) module minimizes catastrophic forgetting by 19% by reducing intra-class variance. APF is computationally inexpensive and can be used with any two-stage detector. The introduced Cosine Margin Cross-Entropy loss increases the angular margin between overlapping classes reducing class confusion by 10%.

References

[1] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. LSTD: A Low-Shot Transfer Detector For Object Detection. In *AAAI*, pages 2836–2843, 2018.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset For Semantic Urban Scene Understanding. In *CVPR*, 2016.

[3] Wentao Cui and Yuhong Guo. Parameterless Transductive Feature Re-representation For Few-Shot Learning. In *ICML*, 2021.

[4] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, pages 248–255, 2009.

[5] 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.

[6] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-Shot Object Detection With Attention-RPN And Multi-Relation Detector. In *CVPR*, 2020.

[7] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized Few-Shot Object Detection Without Forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4527–4536, June 2021.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning For Fast Adaptation Of Deep Networks. In *ICML*, 2017.

[9] Spyros Gidaris and Nikos Komodakis. Dynamic Few-Shot Visual Learning Without Forgetting. In *CVPR*, 2018.

[10] Ross B. Girshick. Fast R-CNN. *ICCV*, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning For Image Recognition. In *CVPR*, 2016.

[12] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot Object Detection Via Feature Reweighting. In *ICCV*, 2019.

[13] Leonid Karlinsky, Joseph Shtok, Sivan Harary, Eli Schwartz, Amit Aides, Rogerio Feris, Raja Giryes, and Alex M. Bronstein. RepMet: Representative-Based Metric Learning For Classification And Few-Shot Object Detection. In *CVPR*, 2019.

[14] Aoxue Li, Weiran Huang, Xu Lan, Jiashi Feng, Zhenguo Li, and Liwei Wang. Boosting Few-Shot Learning With Adaptive Margin Loss. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] Bohao Li, Boyu Yang, Chang Liu, Feng Liu, Rongrong Ji, and Qixiang Ye. Beyond Max-Margin: Class Margin Equilibrium For Few-Shot Object Detection. In *CVPR*, June 2021.

[16] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature Pyramid Networks For Object Detection. In *CVPR*, pages 936–944, 2017.

[17] 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.

[18] 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*, 2020.

[19] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-Margin Softmax Loss For Convolutional Neural Networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 507–516. JMLR.org, 2016.

[20] Anay Majee, Kshitij Agrawal, and Anbumani Subramanian. Few-Shot Learning For Road Object Detection. In *AAAI Workshop on Meta-Learning and MetaDL Challenge*, volume 140, pages 115–126, 2021.

[21] Pascal Mettes, Elise van der Pol, and Cees G M Snoek. Hyperspherical Prototype Networks. In *Advances in Neural Information Processing Systems*, 2019.

[22] Alex Nichol, Joshua Achiam, and John Schulman. On First-Order Meta-Learning Algorithms. *ArXiv*, abs/1803.02999, 2018.

[23] Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman H. Khan, and Fahad Shahbaz Khan. Orthogonal Projection Loss. 2021.

[24] Joseph Redmon and Ali Farhadi. YOLO9000: Better, Faster, Stronger. In *CVPR*, July 2017.

[25] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster R-CNN: Towards Real-Time Object Detection With Region Proposal Networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2015.

[26] Victor Garcia Satorras and Joan Bruna Estrach. Few-Shot Learning With Graph Neural Networks. In *ICLR*, 2018.

[27] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical Networks For Few-shot Learning. In *NeurIPS*, pages 4077–4087, 2017.

[28] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. FSCE: Few-Shot Object Detection Via Contrastive Proposal Encoding. In *CVPR*, June 2021.

[29] 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 *CVPR*, June 2018.

[30] 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.

[31] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching Networks For One Shot Learning. In *NeurIPS*, 2016.

[32] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly Simple Few-Shot Object Detection. In *ICML*, 2020.

[33] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-Learning To Detect Rare Objects. In *ICCV*, 2019.

[34] Yang Xiao and Renaud Marlet. Few-Shot Object Detection And Viewpoint Estimation For Objects In The Wild. In *ECCV*, 2020.

[35] Dongxue Xu and Qijun Zhao. Contrapositive Margin Softmax Loss For Face Verification. ICRCA '18, New York, NY, USA, 2018. Association for Computing Machinery.

[36] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. Meta R-CNN: Towards General Solver For Instance-Level Low-Shot Learning. In *CVPR*, pages 9577–9586, 2019.

[37] Gongjie Zhang, Kaiwen Cui, Rongliang Wu, Shijian Lu, and Yonghong Tian. PNPDet: Efficient Few-Shot Detection Without Forgetting Via Plug-And-Play Sub-Networks. In *WACV*, pages 3823–3832, 2021.

[38] Lu Zhang, Shuigeng Zhou, Jihong Guan, and Ji Zhang. Accurate Few-Shot Object Detection With Support-Query Mutual Guidance And Hybrid Loss. In *CVPR*, June 2021.

[39] Shan Zhang, Dawei Luo, Lei Wang, and Piotr Koniusz. Few-Shot Object Detection By Second-order Pooling. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

[40] Chenchen Zhu, Fangyi Chen, Uzair Ahmed, Zhiqiang Shen, and Marios Savvides. Semantic Relation Reasoning For Shot-Stable Few-Shot Object Detection. In *CVPR*, June 2021.