Meta-Guided Metric Learner for Overcoming Class Confusion in Few-Shot Road Object Detection

Anay Majee **Anbumani Subramanian**
Intel Corporation
firstname.lastname@intel.com

Kshitij Agrawal
Bundl Technologies
kshitij.agrawal@swiggy.in

Abstract

Localization and recognition of less-occurring road objects have been a challenge in autonomous driving applications due to the scarcity of data samples. Few-Shot Object Detection (FSOD) techniques extend the knowledge from existing object classes to learn novel road objects given few training examples. Popular techniques in FSOD adopt either meta or metric learning techniques which are prone to *class confusion* between visually similar objects and *catastrophic forgetting* of already learnt classes. In this work, we introduce a novel Meta-Guided Metric Learner (MGML) which couples the benefits of both meta and metric learners to overcome class confusion in FSOD. We also re-weight the features of the novel classes higher than the base classes through a novel Squeeze and Excite module and encourage the learning of truly discriminative features by applying an Orthogonality Constraint to the meta learner. Our method outperforms existing approaches in FSOD on the India Driving Dataset (IDD) by upto 11 mAP points while suffering from the least class confusion of 20% given only 10 examples of each novel road object. We further show similar improvements on the few-shot splits of PASCAL VOC dataset where we outperform state-of-the-art approaches by upto 5.8 mAP points accross all splits.

1 Introduction

Few-Shot Learning is the ability of Machine Learning models to learn novel concepts from limited training samples [2]. This form of learning has demonstrated its potential to alleviate the requirement of large-scale annotated datasets [5, 18] during model training, which is cumbersome and expensive to obtain. It also has significant importance in real-world scenarios such as autonomous navigation in unconstrained environments to detect less-occurring road objects from few-shot data samples. Unlike standard driving datasets [9, 4], India Driving Dataset (IDD) [29] exhibits a real-world class-imbalanced setting and contains a set of object categories with very few annotated samples [19] as shown in Figure 1(a) - *street carts*, *water tankers*, *tractors* and *excavators*. Approaches tasked to learn from such real-world datasets perform poorly on the less-occurring (*few-shot*) classes.

Recent developments in machine learning research have shown commendable progress in few-shot learning by extending the capability of existing models trained on large-scale datasets to adapt to sparse data. Although these models show exemplary performance for image recognition [8, 10, 16] tasks, Few-Shot Object Detection (FSOD) emerges as a relatively unexplored and complex field as it encompasses both localization and recognition tasks. Traditional approaches in FSOD adopt meta-learning [35, 34] which decomposes the few-shot learning task into multiple subtasks (episodes) and aggregates their learnings through a global objective function [8, 30]. Recent State-of-The-Art (SoTA) approaches have adopted a simpler strategy - metric learning [32, 27], which rapidly adapts to newly introduced (novel) classes by learning discriminative class boundaries between object classes.

Machine Learning for Autonomous Driving Workshop at the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Street Cart Excavator Water Tanker Tractor

(a) Novel road objects in the open-set of the India Driving Dataset (IDD-OS).

(i) (ii)

(b) Challenges in Few-Shot Road Object Detection.

Figure 1: Examples of (a) novel road objects in the Open-Set of IDD and (b) results from existing Few-Shot Object Detection technique FsDet [32] showing the key challenges in few-shot road object Detection. Regions marked in red show (i) class confusion where the *Excavator* is misclassified as an *Autorickshaw* and (i) catastrophic forgetting, where already learnt (base) classes are lost after few-shot adaptation.

Despite the recent successes, SoTA meta and metric learners suffer from *catastrophic forgetting* and *class confusion* as shown in Figure 1(b). This results in loss of information for the already learnt (base) classes and poor performance on the newly identified (novel) classes. We adopt the problem of FSOD in a real-world class-imbalanced setting to detect less-occurring road objects given a few training samples and significantly reduce the impact of class confusion on the performance of base and novel road objects.

Unlike traditional approaches in FSOD, we propose a novel Meta-Guided Metric Learning (MGML) strategy which learns class-specific feature sets through a meta learner and guides a metric learner to eliminate overlapping features between object classes. From our experiments, we notice that a large portion of low-level features is shared between the base and novel classes which when eliminated by the metric learner renders the model ineffective against intra-class variance and inter-class bias. To handle this pitfall the MGML approach introduces a Split-and-Excite module which re-weights the contribution of novel class features significantly higher than the base classes in the predictor head of the few-shot detector. We also apply an orthogonality constraint in the meta learner to encourage the learning of highly discriminative feature sets for each road object. Unlike existing approaches in FSOD which demonstrate their performance on cannonical datasets like PASCAL VOC [6] we adopt the few-shot splits in the challenging IDD-Detection dataset [19] which represents a real-world, class-imbalanced setting. We show that our proposed method overcomes the large inter-class bias and intra-class variance in IDD, and suffers from the least class confusion. The main contributions of our approach can be summarized as:

- We demonstrate that class confusion in FSOD can be overcome with Meta-Guided Metric Learning (MGML) approach which combines both meta and metric learning objectives.
- We demonstrate the learning of truly discriminative class-specific features during model training, by applying an Orthogonality Constraint (OC) and re-weighting the constribution of novel class features relatively higher than the base classes through the Split and Excite (SE) module.

- Through our approach, we demonstrate a performance improvement of upto 11 mAP points while sufferering from the least class confusion of 20.12% on the open-set of the India Driving Dataset (IDD) and demonstrate similar improvements on standard FSOD benchmarks like PASCAL VOC.

2 Related Work

2.1 Few-Shot Learning

Learning algorithms in few-shot learning can be divided broadly into two categories : *Metric Learning* and *Meta-Learning*. Metric learners [30, 26, 28] learn generalizable feature representations from few-shot data, which are used to make predictions on novel tasks. A characteristic property of this class of algorithms is the use of distance / similarity metrices like cosine-similarity [2, 30], euclidean distance [26] and graph distance [25] to adapt to novel classes. Meta-Learners [8, 24] differ from metric learners by the mechanism of encoding the knowledge from few-shot data and propagating it to novel classes. Meta-learners can be further classified into memory based [22], model based [24] and optimization based [8, 20] techniques. Recent works [3, 31] adopt an architecture that combines both metric and meta learning techniques to adapt to novel classes. Our work demonstrates the effectiveness of this technique for object detection task.

2.2 Few-Shot Object Detection

Traditional FSOD techniques [1] adopt transfer learning to adapt to novel classes but suffer from model overfitting and catastrophic forgetting. Metric learning techniques [13, 32, 36] use distance metrices to rapidly adapt to novel classes. FsDet [32] adopts a cosine-similarity based classifier, while PNPDet [36] decouples the base and novel class predictors and learns a cosine-similarity based loss function to reduce model overfitting and class confusion. Another promising direction in FSOD is the use of meta-learning techniques in conjunction with standard object detectors. Techniques like Meta-Reweight [12], Meta-RCNN [35], CME [15] and Add-Info [34] adopt this technique to learn class-specific feature sets to differentiate between base and novel class features. Fan *et al.* [7] learns an Attention-RPN along with a relation network [28] to learn truly discriminative class-specific features to guide the predictor head of the object detector. Modern techniques use vision transformers [37] and contrastive learning [27] to improve performance on novel classes. Li *et al.* [17] combines meta and metric learning techniques by adopting a pearson's distance based metric learner alongside the meta-learner.

FSOD has been recently applied in the context of autonomous driving in [19] to detect less-occuring road objects. The authors in [19] have identified *class confusion* and *catastrophic forgetting* as dominant roadblocks in achieving SoTA performance for road object detection. Our work adopts this problem definition and shows that a combined meta and metric learner can overcome the issue of class confusion while improving performance on novel classes.

3 Method

3.1 Problem Definition

We define a proposal-based few-shot detector $h(I, \theta)$, where I refers to the input data to $h(I, \theta)$ and θ represents the model parameters. We follow the definition of a meta learner as in [19] and train $h(I, \theta)$ in two distinct stages - *base training* and *few-shot adaptation*. $h(I, \theta)$ adopts an episodic [30, 8] training strategy where each episode samples a subset of N classes from the input dataset D (D_{base} during base training and $D_{base} \cup D_{novel}$ during few-shot adaptation) with K examples per class, referred to as *support set* and Q examples ($Q > K$) from D containing N classes, referred to as *query set*. The objective of the few-shot learner $h(I, \theta)$ is to learn generalizable features from abundant training samples in D_{base} during base training and rapidly adapt to novel classes in $D_{base} \cup D_{novel}$ during the few-shot adaptation stage given only K examples for each class.

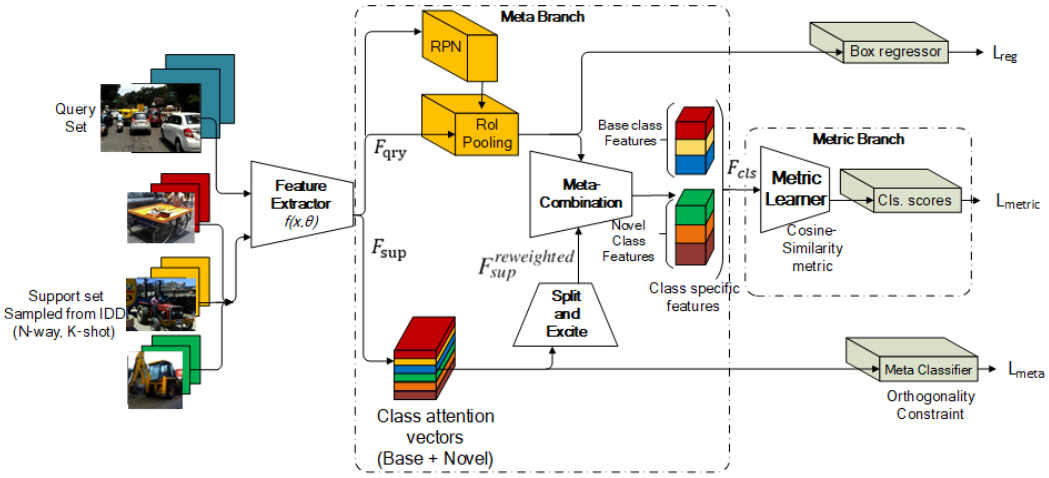

Figure 2: **Architecture of the proposed Meta-Guided Metric Learner (MGML) :** Our approach employs a meta learner which guides a metric learner to learn truly discriminative features from the input dataset to adapt to novel classes.

3.2 Meta Guided Metric Learner

In this work, we introduce a novel Meta-Guided Metric Learner (MGML) which promotes knowledge retention in base classes by learning class-specific feature representations in $D_{base} \cup D_{novel}$ and discriminates between classes by increasing the angular separation between feature clusters.

Unlike traditional meta or metric learning architectures our proposed MGML amalgamates the benefits of both learning strategies to significantly reduce class confusion without any further impact on catastrophic forgetting. As shown in Figure 2, $h(I, \theta)$ produces two sets of features, F_{sup} and F_{qry} from the support and query sets respectively. The features in F_{qry} are class-agnostic as they contain feature information from multiple object classes from the same scene, whereas the features in F_{sup} are class-specfic as each input image carries information of a single object class. The MGML architecture proceeds by decomposing $h(I, \theta)$ into two sequential branches - *meta branch* and *metric branch*. The meta branch learns a set of attention vectors for each class (class-attentive vectors) in the support set. It further follows the Meta-combination module described in [34] to produce attentive feature sets (F_{cls}) for each object in the query set by channel-wise multiplication of the class attentive vectors (F_{sup}) with the class-agnostic features (F_{qry}). Our experiments show significant overlaps among feature sets in F_{cls} which can be attributed to overlapping features in F_{sup}. We mitigate this problem by applying a novel Orthogonality Constraint (OC) described in section 3.2.1.

The metric branch learns distinguishable feature representations for each class by maximizing the class boundaries through a non-linear similarity metric [10]. In this work we adopt a cosine similarity metric as in [32] to minimize the similarity between class-specific feature sets F_{cls} learnt by the meta branch through a metric loss L_{metric}. Despite a significant reduction in confusion, introduction of a metric learner results in a drop in novel class performance especially for classes with strong visual similarities with the base classes. We can attribute this to the elimination of distinguishable features for the confusing novel class objects by the metric learner. To mitigate this pitfall we re-weight the contribution of the novel class attentive vectors significantly higher than the base classes through the novel Split and Excite (SE) module. We describe this in detail in section 3.2.2.

The combined effect of the meta and metric learning objectives along with the SE and OC modules demonstrates significant reduction in class confusion while boosting the performance on novel classes in road object detection tasks [19].

3.2.1 Orthogonality Constraint

For accurate classification of the class specific features in F_{cls} the model $h(I, \theta)$ must learn the most discriminative feature set F_{sup} which uniquely identifies each road object in the input dataset. While standard Cross-Entropy (CE) loss function reduces the likelihood of features belonging to the same

Figure 3: **Qualitative results from the few-shot India Driving Dataset:** We contrast the performance of MGML against FsDet, for novel classes in the IDD-OS split for the 10-shot setting. FsDet suffers from extreme catastrophic forgetting and is unable to adapt to large intra-class and inter-class variations in IDD which are overcome by MGML.

class to be closer in the feature space it does not ensure sufficient angular separation between features from different classes. This is important in few-shot road object detection due to the sparse feature sets learnt from few-shot data and large visual similarity between road objects. Ranasinghe *et al.*[21] imposes orthogonality in feature space for the classification task. We apply a modified orthogonality constraint more suited to few-shot detection task.

The support set in each training episode consists of K examples from N classes in the input dataset D. Each example $\{x_i, y_i\}_{i=1}^{K} \in D$ generates a class attentive vector $F_{sup_{x_i}} = f(x_i, \theta|y_i)$ where x_i and y_i represents the input image and ground truth label and, f is the feature extractor in the meta branch. The orthogonality constraint L_{oc} maximizes the angular separation between vectors from dissimilar classes and minimizes the separation between similar ones. The computation of L_{oc} is described in (1) where the angular separation between vectors is calculated using a cosine-similarity operator.

$$L_{oc} = \sum_{\substack{i,j \in (N \times K) \\ y_i = y_j, y_i \neq \text{BG}}} 1 - cos(F_{sup_{x_i}}, F_{sup_{x_j}}) + \sum_{\substack{i,j \in (N \times K) \\ y_i \neq y_j, y_i \neq \text{BG}}} cos(F_{sup_{x_i}}, F_{sup_{x_j}}) \tag{1}$$

L_{oc} is applied as an additional loss term to the CE loss L_{ce} in the objective function of the meta branch L_{meta} as shown in equation 2. L_{oc} is applied only to the foreground classes as the background (BG) class can potentially contain information from multiple object classes. The hyperparameter α controls the contribution of the orthogonality constraint in L_{meta} and is described in section 5.2.

$$L_{meta} = L_{ce} + \alpha L_{oc} \tag{2}$$

3.2.2 Split and Excite Module

The proposed Split and Excite (SE) module in MGML re-weights the class-specific vectors in F_{sup} pertaining to the novel classes higher with respect to the base classes in the few-shot adaptation stage. This module highlights the sparse features from the novel objects and reduces the chance of feature elimination due to the addition of the metric learning objective. This module can be formulated as three distinct phases. At first, F_{sup} is *split* into base (F_{sup}^{base}) and novel (F_{sup}^{novel}) class vectors. Secondly, the vectors from F_{sup}^{novel} are re-weighted by channel-wise multiplication of a learnable hyper-parameter λ (*excite* phase) as shown in (3) to produce $F_{sup}^{reweighted}$.

$$F_{sup}^{reweighted} = [F_{sup}^{base}, (\lambda * F_{sup}^{novel})] \tag{3}$$

Finally, the base and novel feature vectors are aggregated with the class-agnostic query set features to form the class-attentive feature set F_{cls}. We follow the aggregation process described in [34] to produce the class-specific feature set F_{cls_i} for each class i in the input dataset as shown in equation 4.

$$F_{cls_i} = [F_{qry} \otimes F_{sup}^{reweighted}, F_{qry} - F_{sup}, F_{qry}] \tag{4}$$

Table 1: **Results on few-shot splits of the India Driving Dataset (IDD):** Few-Shot object detection performance on novel classes in IDD-10 and IDD-OS splits from IDD for 5 and 10-shot settings.

Method	Meta/Metic Learner	**IDD-OS (Open-Set)**		**IDD-10 (Split-1)**		**IDD-10 (Split-2)**	
		K=5	10	5	10	5	10
Meta-RCNN [35]	Meta	4.3	6.4	5.7	7.8	7.4	6.7
Add-Info [34]	Meta	18.2	28.8	5.2	10.0	7.7	9.5
FsDet w/ cos [32]	Metric	23.6	39.8	13.1	**22.1**	14.8	**22.8**
Ours (MGML + SE + OC)	Meta + Metric	**28.0**	**48.0**	**15.1**	17.2	**15.2**	18.6

Method	Meta[1] Learner	Metric Learner	SE Module (λ=2.0)	Orthogonality Constraint	mAP_{base}	mAP_{novel}
Add-Info	✓				37.1	28.8
FsDet w/ cos		✓			**47.4**	37.0
MGML (ours)	✓	✓			38.0	40.0
	✓	✓	✓		38.0	45.4
	✓	✓		✓	41.0	46.1
	✓	✓	✓	✓	41.5	**48.0**

Table 2: **Components of the proposed MGML architecture:** Performance comparison among various variants of the MGML architecture shows that the combined meta and metric learner along with the Split-and-Excite (SE) module and Orthogonality Constraint (OC) demonstrates the best novel class performance on the IDD-OS split in the 10-shot setting.

Through our experiments, we empirically show that such a formulation helps to further boost the performance on novel classes without any degradation in the base classes. A more detailed architecture of the SE module has been provided in the appendix section.

3.2.3 Training Methodology

As described in section 3.1, $h(I, \theta)$ is trained in two stages. During the base training stage $h(I, \theta)$ is trained on abundant samples in D_{base} till convergence by adopting the meta training strategy in [35] and applying the orthogonality constraint (L_{meta}) to the meta branch. We use the loss functions in [23] comprising of a binary Cross-Entropy (CE) 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}. During the few-shot adaptation stage, we introduce the metric branch into $h(I, \theta)$ and apply the SE and OC modules in the meta branch to adapt to K-shot data in $D_{base} \cup D_{novel}$. The box classification loss L_{cls} is replaced with a combined meta loss L_{meta} and a cosine similarity penalty L_{metric} as described in equation 5.

$$L = (L_{meta} + L_{metric}) + L_{reg} + L_{rpn} \tag{5}$$

4 Experiments

In this section, we describe our experimental setup and benchmark the performance of our proposed MGML technique on two few-shot object detection benchmarks - India Driving Dataset [19] and PASCAL-VOC dataset [12]. For all our experiments we report the Mean Average Precision (mAP_{50}) at 50% Intersection over Union (IoU) [6], which is a standardized metric for evaluating object detection performance.

4.1 Datasets

India Driving Dataset (IDD) [29] consists of 15 object classes (in the detection dataset), representing traffic scenes on Indian roads. For the few-shot tasks we adopt the benchmark splits in [19]- IDD-10 and IDD-OS, which represents a real-world class imbalanced setting.

IDD-10 consists of 10 representative classes from IDD which are divided into 7 base classes and 3 novel classes. Based on the choice of novel classes we consider two data splits, referred to as split-1 (*bicycle*, *bus* and *truck* as novel classes) and split-2 (*autorickshaw*, *motorcycle* and *truck* as novel classes).

[1] Adapted from Xiao *et al.* [34].

Table 3: 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. Results are averaged over 10 runs. The symbol $-$ indicates that results were not published by the authors.

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 [35]	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 [36]	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 [27]	Metric	FRCN-R101	**28.2**	46.2	54.1	16.5	35.9	45.3	22.2	45.4	49.4
ours (MGML + SE + OC)	Meta + Metric	FRCN-R101	27.3	**51.4**	**58.0**	**22.3**	**37.9**	**45.6**	**22.7**	**47.1**	**51.2**

IDD-OS consists of 14 classes, with 10 base classes and 4 novel classes. The 4 novel classes (*Street cart*, *Tractor*, *Water tanker* and *Excavator*) were generated by expanding on the *vehicle fallback* category in IDD and represents the open world deployment setting.

We use the complete *train* set of IDD for base training and sample N-way K-shot (*K=5, 10*) episodes during few-shot adaptation. We use the *val* set of IDD for evaluation.

PASCAL-VOC [6] consists of 20 object classes, from which 15 are considered as base and 5 are considered as novel classes. The 5 novel classes are randomly chosen to form 3 representative category splits. We follow the data splits from Meta-Reweight [12] and evaluate our methods on novel split-1 (*bird*, *bus*, *cow*, *motorbike* and *sofa*), novel split-2 (*aeroplane*, *bottle*, *cow*, *horse* and *sofa*) and novel split-3 (*boat*, *cat*, *motorbike*, *sheep* and *sofa*) for *K=1, 5 and 10* shot settings. For training we use the complete *trainval* split from PASCAL-VOC07+12 datasets and *test* split of PASCAL-VOC 2007 for evaluation.

4.2 Implementation Details

The MGML architecture is based on the proposal-based Faster-RCNN [23] object detector with a ResNet-101 [11] backbone. Input to the network consists of a batch of 4 query images resized to 600 x 800 pixels and *K*-shot support images from N selected classes in $D_{base} \cup D_{novel}$, resized to 224 x 224 pixels. We consider images with object sizes greater than 100 x 100 pixels for the support set to ensure the quality of extracted features. We follow the base training procedure in [34] to train our model for till convergence (20 epochs). During the few-shot adaptation stage, we train our model for 12 epochs with a constant learning rate of 1 x 10^{-3}. We use an Adam [14] optimizer and momentum value of 0.9. The hyperparameter λ is introduced in the few-shot adaptation stage and α is applied in both stages of model training. The values of α and λ are chosen through ablation experiments in section 5.2. All benchmark experiments are conducted on a single GPU with 12GB memory.

4.3 Results on India Driving Dataset

We benchmark our MGML approach against the FSOD benchmark on IDD as in [19]. Table 1 records the performance of our MGML approach on IDD-10 and IDD-OS splits for 5 and 10 shot settings. We benchmark our approach against both meta and metric learning approaches in [19]. Results show that our approach (MGML + SE + OC) outperforms SoTA approaches on the IDD-OS split by large margins, upto 30% (11 mAP points) across 5 and 10-shot settings. Such results establish the superiority and robustness of our approach in detecting less-occuring road objects in class imbalanced driving environments. On the IDD-10 split, our MGML approach outperforms SoTA meta learning approaches (Add-Info) by upto 9.9 mAP points but under-perform against metric learning approaches. We observe that the MGML does not achieve high performance gains for higher values of K (10-shot) in IDD-10. This can be attributed to the large intra-class variance among road objects in IDD.

Figure 3 demonstrates the qualitative results from our approach against the SoTA metric learner (FsDet) on the IDD-OS split for the 10-shot setting. We demonstrate the robustness of our MGML approach against pitfalls in SoTA FSOD techniques such as large intra-class bias (Figure 3(a),3(b)), catastrophic forgetting (Figure 3(c)) and ineffectiveness against occlusions (Figure 3(b)).

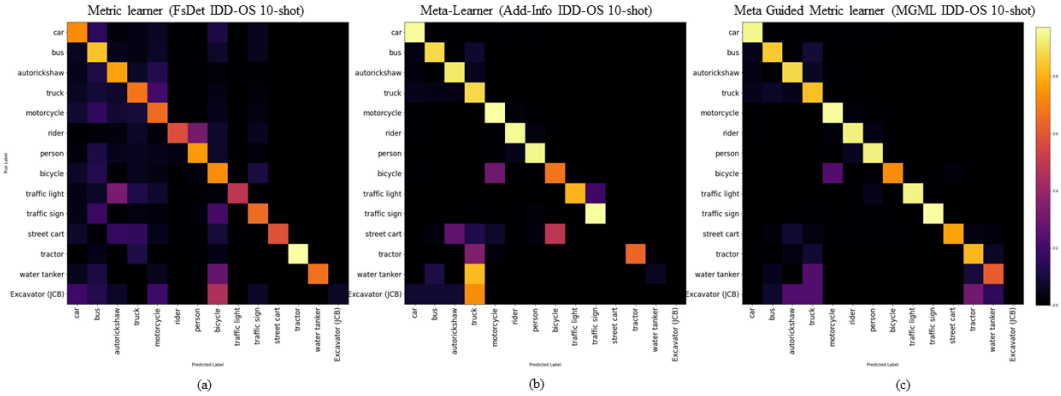

Figure 4: Confusion matrix plotted for class prediction results from three categories of few-shot object detection networks trained on the IDD-OS split for the 10-shot setting. Metric learners (a) show large confusion (upto 40%) between base and novel classes. Meta Learners (b) show better retention in performance of base classes but show large confusion among novel classes. Meta-Guided Metric Learner (c) shows least class confusion of 20.1% with better retention of both base and novel classes.

4.4 Results on PASCAL VOC dataset

We benchmark the MGML approach against SoTA approaches on the few-shot datasplits of the PASCAL VOC dataset as in [27]. Following the datasplits in section 4.1 we conduct our experiments on K=1,5, 10 settings and summarize the results in Table 3. We show that the MGML approach outperforms SoTA approaches on almost all three splits. The maximum improvement was observed in split-1 for the 10-shot setting where MGML approach the SoTA approach (FSCE) by 7.2% (3.9 mAP points). However, slight degradation was observed for

5 Ablations

5.1 Components of Our Proposed MGML Architecture

The MGML architecture can be decomposed into three major components. Table 2 demonstrates the contribution of each component on the base and novel class performance. At first, we combine both meta and metric learners into a single unified architecture where the class specific features learnt by the meta learner guides the metric learner. Secondly, we introduce the SE module which re-weights the novel class attention vectors higher than base classes to reduce chances of feature elimination during few-shot adaptation. Finally, we apply a novel Orthogonality Constrant on the meta branch to learn the most distinguishable feature vector for each class. The combined effect of all the three components demonstrates the best novel class performance on the IDD-OS split in the 10-shot setting but continues to suffer from base class forgetting due to large inter-class bias among road objects.

Table 4: Ablation for key hyper-parameters (λ and α) in the MGML approach showing their impact on base and novel class performance for the IDD-OS (10-shot setting) split. The value of each hyper-parameter chosen for the MGML approach is underlined and associated performance scores are indicated in **bold**.

Parameter	Value	mAP_{base}	mAP_{novel}
λ (α = 0.0)	1.0	37.9	40.0
	1.5	38.0	42.8
	2.0	**38.1**	**45.0**
	2.5	37.0	40.9
α ($\lambda = 2.0$)	0.05	40.5	46.8
	0.1	40.8	47.4
	0.5	**41.0**	**47.9**
	1.0	40.9	47.0
	2.0	41.0	45.7

5.2 Hyper-parameters of The MGML Technique

The formulation of MGML introduces two hyper-parameters λ and α. Initially we choose the value of $\alpha = 0$ and vary the value of λ in the range of 1.0 to 2.5. We observe an increase in novel class performance with increase in λ between 1.0 and 2.0 and a steep reduction thereafter. Thus we chose $\lambda = 2.0$ for all our experiments. We then fixate the value of λ at 2.0 and vary the value of α in the range of [0.05, 2.0]. We observe a small improvement in novel class performance in the range of 0.05 to 0.5. We also observe a significant boost in base class performance with the introduction of the OC constraint. Based on the results in table 4 we choose the value of α as 0.5 for all our benchmark experiments.

5.3 Class Confusion Between Base and Novel Classes

Class confusion stands out as a prominent issue in few-shot object detection in the context of autonomous driving as road objects share a large number of visual features [19] among themselves. Figure 4 compares class confusion of our approach (MGML) with popular FSOD architectures (FsDet and Add-Info) through confusion matrices. FsDet (metric learner) shows the highest confusion (43.9 %) between both base and novel classes. On the other hand, Add-Info (Meta-Learner) shows reduced confusion among base classes but elevated confusion among novel classes (average confusion of 31.05 %). Finally, MGML shows the least confusion between both base and novel classes (20.12 %). We attribute the contribution of the meta branch to the reduction in confusion among base classes and the metric branch for the improvements on the novel classes. Objects with strong visual dissimilarity with existing classes in the dataset like *excavator* continue to suffer from class confusion resulting from the lack of samples in the few-shot setting.

6 Conclusion

In this work, we introduced a novel FSOD technique - Meta-Guided Metric Learner (MGML) to overcome the impeding effect of class confusion in detecting less-occurring road objects in driving environments. Unlike existing approaches in FSOD we employ both meta and metric learning objectives in an unified proposal-based architecture. The introduced Orthogonality Constraint and Split and Excite module ensures the learning of discriminative feature sets to overcome inter-class variance and intra-class bias among road objects. Our approach achieves SoTA performance on the open-set split of the challenging India Driving Dataset by demonstrating upto 30% improvement in novel class performance over existing methods in a real-world class-imbalanced setting. Alongside improvements in absolute performance, the MGML approach suffers the least class confusion of 20.12% among SoTA FSOD approaches. Catastrophic forgetting of base classes continues to be an impeding issue in few-shot road object detection and will be addressed in future research.

7 Acknowledgements

The authors would like to thank Pravin Chandran, Ashutosh Agarwal and Raghavendra Bhat for their valuable comments and discussions during this work.

References

[1] Hao Chen, Yali Wang, Guoyou Wang, and Yu Qiao. LSTD: A low-shot transfer detector for object detection. In *Proc. of the Thirty-Second AAAI Conf. on Artificial Intelligence*, pages 2836–2843, 2018.

[2] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *Intl. Conf. on Learning Representations*, 2019.

[3] Yu Cheng, Mo Yu, Xiaoxiao Guo, and Bowen Zhou. Few-shot learning with meta metric learners. *ArXiv*, abs/1901.09890, 2019.

[4] 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 *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

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

[7] Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai. Few-shot object detection with Attention-RPN and Multi-Relation detector. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for fast adaptation of deep networks. In *Proc. of the 34th Intl. Conf. on Machine Learning*, 2017.

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? the KITTI Vision Benchmark Suite. In *Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[10] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for Image Recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *IEEE Intl. Conf. on Computer Vision (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 *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd Intl. Conf. on Learning Representations, ICLR*, 2015.

[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] Wenbin Li, Lei Wang, Jinglin Xu, Jing Huo, Yang Gao, and Jiebo Luo. Revisiting local descriptor based image-to-class measure for few-shot learning. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] Yuewen Li, Wenquan Feng, Shuchang Lyu, Qi Zhao, and Xuliang Li. Mm-fsod: Meta and metric integrated few-shot object detection. 12 2020.

[18] 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 *Computer Vision – ECCV 2014*, pages 740–755, 2014.

[19] 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 of *Proceedings of Machine Learning Research*, pages 115–126, 2021.

[20] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *ArXiv*, abs/1803.02999, 2018.

[21] Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman H. Khan, and Fahad Shahbaz Khan. Orthogonal projection loss. 2021.

[22] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *5th Intl. Conf. on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[23] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015.

[24] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *Proc. of The 33rd Intl. Conf. on Machine Learning*, volume 48, pages 1842–1850, 2016.

[25] Victor Garcia Satorras and Joan Bruna Estrach. Few-shot learning with graph neural networks. In *Intl. Conf. on Learning Representations*, 2018.

[26] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[27] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. Fsce: Few-shot object detection via contrastive proposal encoding. In *Proc. of the IEEE conf. on computer vision and pattern recognition (CVPR)*, June 2021.

[28] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to Compare: Relation network for few-shot learning. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[29] 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 *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, pages 1743–1751, 2019.

[30] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 2016.

[31] Duo Wang, Yu Cheng, Mo Yu, Xiaoxiao Guo, and Tao Zhang. A hybrid approach with optimization-based and metric-based meta-learner for few-shot learning. *Neurocomputing*, 349:202–211, 2019.

[32] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E Gonzalez, and Fisher Yu. Frustratingly simple few-shot object detection. In *Intl. Conf. on Machine Learning (ICML)*, 2020.

[33] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Meta-learning to detect rare objects. In *IEEE Intl. Conf. on Computer Vision (ICCV)*, 2019.

[34] Yang Xiao and Renaud Marlet. Few-Shot object detection and viewpoint estimation for objects in the wild. In *European Conf. on Computer Vision (ECCV)*, 2020.

[35] 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 *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 9577–9586, 2019.

[36] 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 *Proc. of the IEEE/CVF Winter Conf. on Appl. of Computer Vision (WACV)*, pages 3823–3832, 2021.

[37] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, and Shijian Lu. Meta-detr: Few-shot object detection via unified image-level meta-learning, 2021.

A Appendix

A.1 Architecture of The Split and Excite Module

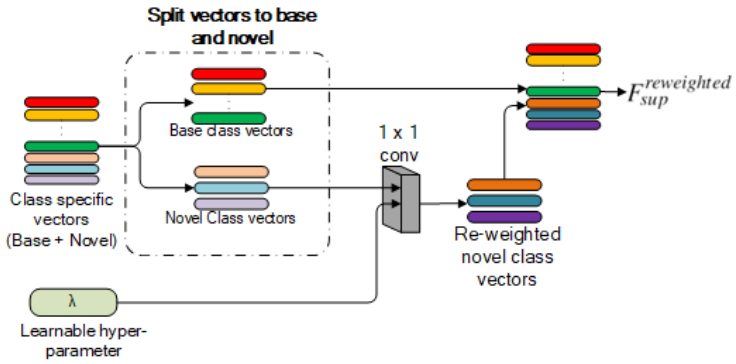

Figure 5: Architecture of the proposed Split and Excite module.

Following the definition of the Split and Excite (SE) module in section 3.2.2 the sub-network can be decomposed into three distinct parts. Figure 5 demonstrates the detailed architecture of the Split and Excite module. The output of the meta branch is a set of class specific attention vectors F_{sup}. At first, we segregate the class attention vectors into base and novel classes. Secondly, we pass the attention vectors belonging to the novel classes and the learnable hyper-parameter λ through a 1×1 convolution layer which re-weights the novel class vectors higher than the base classes, $\lambda > 1$. Finally, we concatenate the re-weighted feature vectors of the novel classes with those of the base classes to produce $F_{sup}^{reweighted}$.

A.2 Additional Qualitative Results on IDD

Figure 6 demonstrates addiitonal qualitative results from the MGML approach on the novel classes in IDD-OS split of the India Driving Dataset (IDD) in the 10-shot setting. The positive results indicates that our method is resistant to the major pitfalls in standard object detection algorithms - occlusions, variational lighting etc. It also points out the reduction in class confusion between co-occuring and visually similar classes such as *street cart* and *bicycle* (visually similar), or *motorcycle* and *rider* (co-occuring). The negative results indicate the existence of class confusion in object classes which have large visual similarity with existing classes such as *Excavator (JCB)*. Such issues will be addressed in future research.

Positive results from novel classes in IDD-OS for 10-shot setting.

Negative predictions on IDD-OS split in 10-shot setting.

Figure 6: Qualitative results from the MGML approach on the few-shot India Driving Dataset (IDD-OS split) for the 10-shot setting. The first three rows in the figure show positive predictions while the final row shows failure cases where we continue to observe class confusion and catastrophic forgetting.