Information Maximization to Overcome Catastrophic Forgetting in Few-Shot Object Detection

Aditeya Baral*
PES University
aditeya.baral@gmail.com

Anay Majee
Intel Corporation
anay.majee@intel.com

Anbumani Subramanian
Intel Corporation
anbumani.subramanian@intel.com

Abstract

*Few-shot object detection encompasses the tasks of localizing and classifying objects in an image provided a limited number of training examples. Recent techniques in this domain suffer from confusion between object classes and demonstrate a tendency to forget the knowledge of already learnt classes, also known as catastrophic forgetting. Our work overcomes the impedance of catastrophic forgetting through an information maximization approach – **I**nformation **M**aximization **N**etwork (**IMNet**) that focuses on learning more descriptive feature representations without overfitting on irrelevant ones while retaining the relevant features from already learnt classes in an input image. Our introduced Cross-Entropy Similarity Loss decreases class confusion by adjusting the embedding space to allow homogeneous classes to have feature representations close to one another and heterogeneous classes to have a high separation between them. We conduct our experiments on the India Driving Dataset (IDD), which demonstrates a real-world setting alongside large class imbalance. Our IMNet architecture outperforms existing meta-learning approaches by 0.2 mAP on the base classes and up to 3 mAP on novel classes of IDD.*

1. Introduction

Convolutional Neural Networks [14] have greatly improved Computer Vision for decades now and help obtain incredibly high performance on various tasks ranging from classification, segmentation and object detection [9, 10, 24]. However, these architectures are dependent on large-scale image datasets [12, 23, 26] to achieve State-of-the-Art performance. However, such datasets may not be readily available as real-world data is expensive to acquire and labour-intensive to annotate.

Current efforts in few-shot and meta-learning attempt to mitigate this problem by adapting to a few training examples per class. Significant progress has been made in this domain, particularly in computer vision tasks like image classification [30, 34], to identify novel instances from limited data. Unlike classification, the joint tasks of localization and recognition, coupled with the few-shot learning setting, render few-shot object detection as an extremely challenging yet relatively unexplored task. This challenge is further exemplified in real-world settings under large variations in structural and environmental conditions [31].

Efforts to tackle few-shot object detection has taken two prominent approaches – Metric and Meta-Learning. Metric learners [28, 35] learn discriminative feature embeddings for each class by adopting a suitable distance/ similarity metric to compute class probabilities. Meta learners [2, 3, 8, 17, 32, 38] train and adapt to new few-shot training tasks by learning an optimization strategy to maximize performance on each task. Experiments show that metric-learners significantly outperform meta-learners when it comes to adapting to few-shot data. However, both meta and metric learners are plagued with impedances like class confusion and catastrophic forgetting [21, 22].

Class confusion [22] occurs when a model is unable to distinguish between object classes owing to its failure in learning discriminative features from few-shot data. In object detection, this manifests when a Region of Interest (RoI) is misclassified as an incorrect class, especially in newly added (novel) classes which are confused to be one or more visually similar classes from the already learnt (base) ones. Catastrophic forgetting [25] is the degradation in performance of the base classes when adapting to novel ones. This occurs mostly in real-world settings where objects share a large volume of intrinsic features resulting in the model overfitting on the few-shot classes. In systems like autonomous driving where vision algorithms are tasked to detect less-occurring road objects [21], the problems of class-confusion and catastrophic forgetting are prominently visible.

The current trend in few-shot object detection (described in section 2) shows that the performance of a few-shot detector improves with its ability to learn more descriptive fea-

*Work done as an intern at Intel.

ture representations [15, 36]. Recent techniques [13] prevent model overfitting by encouraging the model to look beyond the irrelevant features [16] and pick up the distinct characteristics of an object that make it stand out from the rest. The advent of the attention mechanism [1] in natural language processing has opened new doors to the creation of such representations. Similar attention-based techniques have been applied to computer vision tasks [13] which encourage the learning of spatial relationships between feature representations and allow discriminative features to be weighted higher than the overlapping ones.

In this work, we propose a novel meta-learning and information maximization based approach named **I**nformation **M**aximization **N**etwork (**IMNet**) to overcome the catastrophic forgetting. We form descriptive feature representations by computing spatial relationships within the same image as well as between image pairs using self-correlation and co-attention mechanisms respectively. This allows the network to focus on a small subset of discriminative features instead of the wider range of irrelevant ones to distinguish between objects. Additionally, our Cross-Entropy Similarity Loss adjusts these representations to increase the inter-class separation between objects and reduce class confusion.

Existing works on few-shot object detection evaluate their model's performance on canonical benchmark datasets such as the PASCAL-VOC [6] and MS-COCO [19], which do not represent real-world scenarios exhibiting high intra-class variance and inter-class bias. In the real-world, autonomous driving faces these challenges and for this reason, we evaluate the performance of our approach on the India Driving Dataset (IDD) [31].

The main contributions of our work can be summarized as:

- We show that catastrophic forgetting in few-shot object detection can be overcome through a meta-learning based information maximization approach (IMNet) in driving scenes.
- We demonstrate the aggregation of feature representations from visually similar object classes through a Self-Correlation (SCR) module and learning of discriminative class-specific features through a Cross-Correlation (CCR) module.
- Our IMNet approach demonstrates a performance improvement of up to 3 mAP points on real-world driving conditions in the India Driving Dataset (IDD) while retaining performance on already learnt (base) classes.

2. Related Work

2.1. Few-Shot Learning

Existing works in few-shot learning can be classified into two main categories - *meta* and *metric* learning. Metric learners [28, 35] attempt to learn and use an embedding space from the provided base data along with a suitable distance metric to minimize the distance between an object and its corresponding class label. However, metric learners have been observed to suffer from catastrophic forgetting [25], since they tend to easily overfit on the few-shot novel class data. This causes the learner to perform well on the novel classes but lose performance on base classes. Meta learners [2, 3, 17, 32, 38] divide the few-shot task into episodes or tasks and try to maximize performance on each episode by adapting base class learning to novel classes. These learners have been able to mitigate the problem of catastrophic forgetting to a suitable extent, with current approaches trying to experiment with generalizability of the feature space [7].

2.2. Few-Shot Object Detection

Early works in few-shot object detection use a distance-based metric learner to extend knowledge of the learnt and abundant base classes to the novel classes. The authors in [21] compare the performance of metric and meta-learning on the India Driving Dataset (IDD) and observe that metric learning outperforms meta-learning by a significant margin.

Recently, meta-learning has been looked at as an alternative for few-shot object detection. The authors in [15] tackle class-confusion and high intra-class variance in the support using spatial aggregation methods and specific support and query attention components and hypothesise that the distribution of the query can be significantly different from the support.

[36] argues that meta-learning is incapable of learning invariant object characteristics and thus class-specific prototypes are ineffective. They create a universal prototype that learns from all object classes and then refine it to extract class-specific feature representations. [16] tries to improve the learner's generalizability by applying transformations such as warping and removing patches to the support and query images in the training stage, thus forcing the learner to learn transformation invariant characteristics. They show that they are capable of reducing the generalization gap between the fully supervised and few-shot settings. [39] uses a hallucinator network to generate additional training examples in the RoI feature space to handle variations in the test data. They found that hallucination improved performance in extremely low few-shot settings.

The authors in [7] display no drop in base class performance and achieve State-of-the-Art results on the PASCAL-VOC dataset by using a bias-balanced RPN which adapts better to novel class data without forgetting past knowledge.

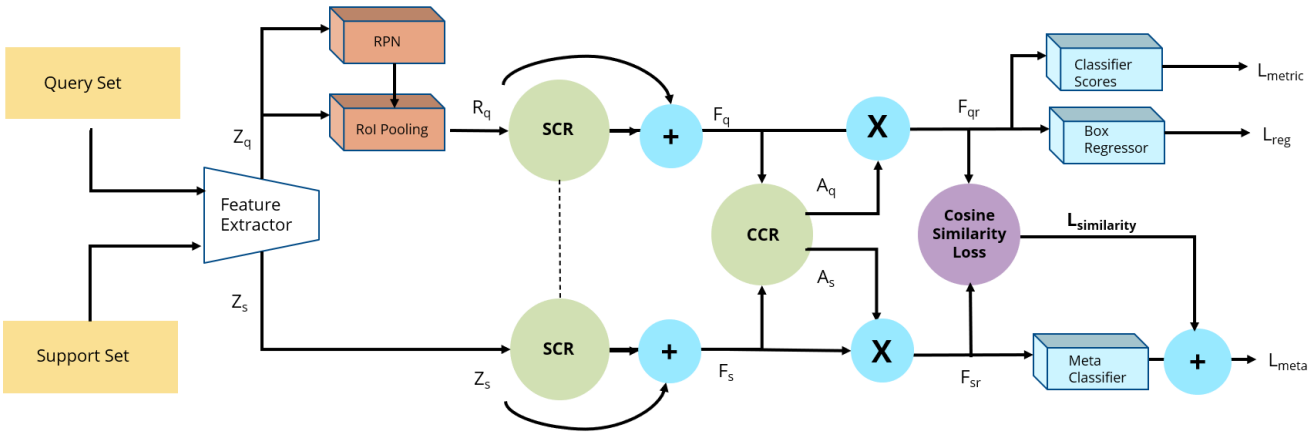

Figure 1. **The architecture of our proposed approach IMNet**: After base training on abundant large scale data, the model is fine-tuned on K-shot examples from base and novel classes. The SCR and CCR are attention mechanisms used to learn more discriminative feature representations and the Cosine Similarity Loss is used to increase the separation between learnt feature representations. The entire architecture has been explained in section 3.2

However, when evaluated on the India Driving Dataset (IDD), their approach was not able to handle the high class imbalance and performed poorly.

2.3. Few-Shot Image Classification

The current trend in few-shot image classification is to learn more descriptive feature representations for images to encode more useful information and suppress irrelevant information which plays no part in distinguishing one image from another. The authors of [4] state that it is often the subtle traits that help discriminate between two objects and that major image features are class irrelevant. They decouple the representation learning into two branches to learn discriminative and irrelevant features independently and then use a decoder network to combine the representations. [20] uses a contrastive learning approach to train two separate encoders to maximize the similarity between features of the same class.

PropNet [27] tackles scenarios with background noise and occlusions by performing localization conditioned classification. The authors of ReNet [13] introduce the concept of relational embeddings and use two attention mechanisms that aggregate information from within the same image as well as across two images to enhance feature representations, which greatly improves performance and helps them achieve State-of-the-Art on benchmark datasets. Our work adopts the attention mechanisms from ReNet for object detection and combines it with the contrastive learning approach to learn highly discriminative feature representations.

3. Method

In this section, we define the Few-Shot Object Detection problem and describe our proposed Information Maximization Network approach.

3.1. Problem Definition

We define a proposal based few-shot object detector $h(I, \theta)$ consisting of a class-agnostic component $f(I, \theta_f)$ and a class-specific component $C(f, \theta_c)$ 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 represent the respective model parameters for the components of the few-shot object detector. We propose a meta learning based training approach consisting of two stages: *base training* and *few-shot adaptation* or *fine-tuning*. An episodic training strategy is followed where each episode samples N classes from D (D_{base} during base-training or from $D_{base} \cup D_{novel}$ during fine-tuning), containing K examples per class, known as the Support Set S and a set of query images, Q, where $Q > K$ from D containing N classes, known as the Query Set.

During base training, $h(I, \theta)$ learns to detect objects given sufficient training examples of base classes (C_{base}) from dataset D_{base}. In the fine-tuning stage, $h(I, \theta)$ is fine-tuned using images in D_{novel} which consists of classes $C_{base} \cup C_{novel}$ with only K instances from N classes such that $|C_{base} \cup C_{novel}| = N$. The objective of $h(I, \theta)$ is to boost performance on novel classes in D_{novel} while retaining (or with minimal degradation of) performance on classes in D_{base}.

3.2. Information Maximization Network

Due to the absence of large-scale training data in few-shot object detection, the feature extractor ($f(I, \theta_f)$) tends to overfit on low-level features from classes with abundant data samples, such as the wheels of a *bicycle* while predicting a novel object like the *street cart*. For most of the novel classes, these features are irrelevant and do not help distinguish the novel from the base classes [13, 16]. As a result, novel classes are seldom confused with base classes. On the other hand, learning a similarity-based loss function helps discriminate between object classes, but leads to the model overfitting on the novel classes resulting in catastrophic forgetting.

We propose an information maximization based approach - Information Maximization Network (IMNet) which adopts a novel meta-learning strategy to learn more descriptive feature representations and help reduce catastrophic forgetting. Figure 1 represents our proposed IMNet architecture, which consists of two major learnable units – the *self-correlational (SCR)* and *cross-correlational (CCR)* units. These units learn feature representations that encode the most discriminatory features of objects while ensuring that the irrelevant features are suppressed in the process.

The IMNet architecture can be decomposed into two parallel branches based on the data they operate on. The first operates on the support images I_s to extract feature information Z_s for each object class. The second operates on the query set I_q containing multiple objects in a single scene. Unlike classification, the latter extracts feature representations R_q, for each Region-of-Interest (RoI) in the image, as predicted by the Region Proposal Network (RPN) in $h(I, \theta)$. The feature extractor (a residual network in IMNet) shares its weights across both the branches to ensure the projection of the support and query features on the same embedding space. The SCR described in 3.3 applies channel-wise attention to each input image by aggregating similar features and structural patterns and produces attentive feature sets F_s and F_q for the support and query set images respectively. The CCR modules described in 3.4 computes the spatial co-attention between F_s and F_q, allowing the query features to learn from the support and reduce the intra-class variance between similar classes in the support and query sets. We also introduce a Cross-Entropy Similarity loss function in 3.5 to jointly train the support and query branches while encouraging the feature alignment between similar classes in both branches. Together we show that our proposed IMNet architecture is effective in reducing catastrophic forgetting while improving performance on novel classes.

3.3. Self-Correlational Attention

The SCR unit transforms the base representations from the feature extractor to allow them to focus on the more relevant features present in an image. This prevents the feature extractor from focusing on larger and common image features which are distributed across objects and produce feature representations that look at only the discriminative features of each object which are sometimes overlooked. These representations are more descriptive and are fed as input to the CCR unit for computing co-attention.

The SCR in IMNet adopts a convolution-based attention module as in [13] with modifications towards the few-shot object detection task.

It receives the base representation $Z_s \in \mathbb{R}^{N \times H \times W \times C}$ from the feature extractor for the support and $R_q \in \mathbb{R}^{P \times H \times W \times C}$ for the query, where P represents the number of RoIs in the query set, H and W are the height and width of each RoI feature vector, C is the number of channels and N is the number of classes in the input dataset. It then computes the self-correlation for each feature representation to produce $F_s \in \mathbb{R}^{N \times H \times W \times C}$ and $F_q \in \mathbb{R}^{P \times H \times W \times C}$. The self-attentive representations are then combined back with the base representations of the support and query respectively to generate the final feature representations from the SCR unit.

3.4. Cross-Correlational Attention

The large variability among object instances between the support and the query sets in open-world settings render meta-learning based approaches [37,38] ineffective towards generalizing to novel objects. Encouraging the detector to learn the relationship between support and query set images [13] has proven to be effective in overcoming this barrier for classification tasks. We adopt a similar strategy for object detection and show its effectiveness towards learning class-specific feature sets attentive to only the most discriminative features for each object class.

The CCR unit transforms the SCR feature representations F_s and F_q by first computing the co-attention between them and then producing co-attention maps. These co-attention maps create a relation between the support and the query and allow the query to align its feature representations with the support such that it can learn feature representations for the pooled RoI based on the discriminatory features present in the support feature representation for each class. This allows the query to identify structural patterns in the pooled RoI corresponding to each support class and can easily distinguish one object from another.

A weighted softmax over the computed co-attention maps produces the final CCR attentive feature representations $A_q \in \mathbb{R}^{P \times H \times W \times C}$ and $A_s \in \mathbb{R}^{N \times H \times W \times C}$. The CCR representations are then aggregated with F_q and F_s respectively and then reduced to produce $F_{qr} \in \mathbb{R}^{P \times S}$ and $F_{sr} \in \mathbb{R}^{N \times S}$, where S is the reduced dimension size of the output CCR feature representation.

3.5. Cross-Entropy Similarity Loss

Although the newly learnt attentive feature representations encode more relevant information about the support and query respectively, a significant overlap exists between the feature clusters as open-world objects share a large volume of low-level features. Earlier research [22] shows the effectiveness of orthogonality based objective functions in improving the discrimination between overlapping clusters resulting in better separability between highly similar object classes.

The Cross-Entropy Similarity loss is applied to the output of the CCR module, between the N classes in the support and all the P RoI features in the query. A pairwise similarity matrix $U \in \mathbb{R}^{N \times P}$ as shown in (1) is constructed to quantify the similarity between each feature pair.

$$U = \frac{F_{sr} \,.\, F_{qr}}{\|F_{sr}\| \|F_{qr}\|} \tag{1}$$

Using ground truth labels $y \in \mathbb{R}^{P}$ as the target, we compute the cross-entropy loss over this similarity matrix for each class in the support (Equation 2), and then aggregate all the losses as shown in 3).

The objective function $l_{similarity}$ maximizes the inter-class distance between heterogeneous pairs while reducing the distance between homogeneous classes. This allows the feature representations to reduce class confusion and easily distinguish between similar-looking objects. The hyperparameter, λ allows us to vary the contribution of $L_{similarity}$ and has been tuned during ablation studies and set to a constant value of 0.25.

$$l_{similarity(i \in N)} = -log \frac{e^{U_i, y_i}}{\sum_{n=1}^{N} e^{U_{i,n}}} \tag{2}$$

$$L_{similarity} = \lambda \sum_{i=1}^{N} l_{similarity(i)} \tag{3}$$

We do not compute the Cross-Entropy Similarity loss on the "background" class to prevent loss of information during the model training stage since the "background" class proposals may contain features belonging to one or more object classes.

3.6. Training Objective

Unlike recent approaches [21, 33] in FSOD our proposed IMNet architecture adopts an end-to-end training strategy. Since we adopt a meta-learning based training procedure alongside a feature attention mechanism, our entire network is trained both during the base training as well as the fine-tuning stage in contrast to recent approaches which only finetune a few specific network layers.

Our model $h(I, \theta)$ adopts a Faster-RCNN [24] based object detector as the backbone for the feature extractor. The base training stage trains the model $h(I, \theta)$ on D_{base} which contains abundant training examples until convergence and thus adopts the meta training strategy proposed in [33]. The object detection objectives are adopted from [24] which comprise of a binary cross-entropy loss used at the Region Proposal Network (RPN) to separate foreground and background objects and obtain L_{rpn}, a cross-entropy loss for the bounding box classifier L_{cls} and a smoothed L1 loss for localizing the deltas of the bounding box L_{reg}. We also apply our Cross-Entropy Similarity loss $L_{similarity}$ to increase separation between feature representations between objects. During the few-shot adaptation stage, the model adapts to K-shot data in $D_{base} \cup D_{novel}$ and the box classification loss is replaced with a combined meta loss L_{meta} which consists of the Cross-Entropy Similarity loss $L_{similarity}$ and a cosine similarity penalty L_{metric} as described in equation 4.

$$L = L_{similarity} + L_{meta} + L_{metric} + L_{reg} \tag{4}$$

4. Experiments

In this section, we describe our experimental setup and evaluate our proposed IMNet approach on standard benchmarks and compare the results with the State-of-the-Art. Ablation studies are also performed to validate the effect of our model's major components. The standard evaluation criterion from [33, 37] is adopted to report the Mean Average Precision (*mAP*) at 50% Intersection Over Union (IoU) for all experiments.

4.1. Dataset

For evaluation, we use the **India Driving Dataset (IDD)** [31] because it represents a real-world class-imbalanced scenario in which the classes exhibit high intra-class variance and inter-class bias. The IDD comprises 15 object classes representing objects seen on Indian roads. The dataset consists of two few-shot data splits [21] – the **IDD-OS** which consists of 10 base classes and 4 novel classes and the **IDD-10** which consists of 7 base classes and 3 randomly chosen (based on the few-shot split) novel classes. We evaluate our proposed approach on the complete validation set of the IDD-OS split for the 10-shot setting.

4.2. Experiment Setup

The IMNet architecture is based on the Faster-RCNN [24] model with a ResNet-101 [11] and Feature Pyramidal Network [18] backbone. The architecture is adapted from MGML [22], which is the current State-of-the-Art on the IDD. The input resolution is set to 1920 x 1080 and we use an input batch size as 1 to the network. As highlighted in 3.6, we train the model until convergence with a learning

rate of 0.001 for both base training as well as few-shot fine-tuning. Base training is performed with a pre-trained ImageNet [5] backbone. Horizontal image flipping, random cropping and other standard data augmentation techniques are applied as well. The optimum local window size for computation of the SCR embedding is found using ablation and is set to 32 and a kernel size of $[3 \times 3 \times 3 \times 3]$ is used for performing 4D convolutions in the CCR module. The hyperparameters chosen for the IMNet model such as the attention units (SCR and CCR), planes (p), lambda (λ) are chosen through ablation experiments, as shown in section 4.4. All results from existing methods have been a reproduction of the algorithm from repositories that are publicly available. All our experiments have been performed on a single GPU with 12 GB of memory.

4.3. Results

Our proposed IMNet approach is benchmarked against State-of-the-Art (SOTA) meta as well as metric learners on the IDD-OS split. Table 1 contrasts the base and novel class performance of IMNet against meta-learning techniques such as the Meta-RCNN [38] and Add-Info [37] and metric learning techniques such as FsDet [33] and FSCE [29]. Our approach outperforms the existing State-of-the-Art meta+metric learner, MGML [22] by 0.2 mAP points on base classes and by 3 mAP points on novel classes respectively. Qualitative results from our approach shown in Figure 2 re-affirms its capability in improving performance on novel classes without any significant degradation in the base classes, thus showing its effectiveness against catastrophic forgetting.

Table 3 show that the CCR component primarily influences the reduction in catastrophic base-class forgetting. However, it achieves this by compromising performance on novel classes. Additionally, we observe that our model can achieve higher mAP scores on all base classes compared to the benchmark. Using only the SCR component, the IMNet architecture can achieve a significant improvement in novel class performance on two object categories – *Water Tanker* and *Tractor* while achieving similar performance on base classes.

4.4. Ablation

In this section, we conduct ablation experiments on the IDD-OS split to qualify the effects of the core components of our proposed IMNet architecture and choose optimum values for the corresponding hyperparameters. All results have been compared with MGML [22] since it is the current State-of-the-Art on the IDD-OS split.

Table 1. **Results on Few-Shot India Driving Dataset**: Few-shot object detection performance on the IDD-OS split in the 10-shot setting. Our Method outperforms existing benchmarks by upto 3 mAP points on the novel classes while retaining the performance on base classes.

Method	**Learner**	mAP_{base}	mAP_{novel}
Meta-RCNN [38]	Meta	24.1	4.3
Add-Info [37]	Meta	36.4	18.2
FSDet w/ cos [33]	Metric	38.2	23.6
FSCE [29]	Metric	38.1	39.1
MGML [22]	Meta + Metric	40.6	44.8
IMNet (ours)	Meta	**40.8**	**47.8**

Table 2. **Ablation on various components of the proposed IMNet approach**: A ✓ under the *Base Training* column indicates the component was included during the base training stage as well. The performance leading to the highest mAP score on novel classes has been mentioned in **bold**.

Method	**SCR**	**CCR**	**Base Training**	mAP_{base}	mAP_{novel}
MGML [22]	-	-	-	40.6	44.8
IMNet (ours)	✓			**40.8**	**47.8**
	✓		✓	42.0	41.5
	✓	✓		39.6	38.8
	✓	✓	✓	42.0	44.0

4.4.1 Ablation on Components of the IMNet Architecture

In this section, we apply individual components of the IMNet architecture in different stages of training and also vary the training procedure to identify the best-suited architecture that overcomes the challenge of catastrophic forgetting. We follow the conclusions from [13] and apply the SCR to both the support and query branches while proposing two different variations of the IMNet architecture.

At first, we adopt a simple fine-tuning strategy to learn the SCR and CCR modules of IMNet. The base training procedure at this stage follows MGML with the CCR and SCR modules applied only during finetuning. Table 2 shows the variation in performance of IMNet when SCR ad CCR modules are applied both separately and jointly during training. Our results show that the application of only the SCR module to both the support and query sets produces the best overall performance.

Secondly, we adopt a joint training strategy and include the SCR and CCR modules even during the base training stage. Table 2 shows the variations in performance during joint training and concludes that base training of either the SCR or CCR modules does not result in any performance gains. Despite low gains in overall performance, joint training shows significant performance improvements when the SCR and CCR modules are both applied in the IMNet architecture (row 2 of Table 2).

Table 3. **Results on Base Classes in Few-Shot Splits of India Driving Dataset**: Few-shot object detection performance (mAP_{50}) on base classes on the IDD-OS split on the 10-shot setting. The results show that our IMNet outperforms existing approaches in developing a resistance for catastrophic forgetting.

Method	SCR	CCR	Autorickshaw	Car	Motorcycle	Person	Rider	Traffic light	Traffic sign	Truck	Bicycle	Bus	mAP_{base}
MGML [22]	-	-	54.7	52.2	50.5	35.3	38.2	9.6	22.9	49.6	33.8	59.4	40.6
IMNet (ours)	✓		53.6	52.0	50.5	34.7	37.5	**13.4**	**23.2**	49.6	35.5	58.3	40.8
	✓	✓	**57.5**	**52.9**	**51.4**	**36.1**	**39.7**	11.5	22.2	**52.7**	**37.6**	**58.9**	**42.0**

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

Figure 2. **Comparison of qualitative results from the few-shot India Driving Dataset between MGML and IMNet**: The rows in the figure represent predictions made by MGML [22] and our IMNet approach respectively. We observe that our proposed network outperforms MGML on novel classes of IDD given only 10 examples of each class. Classes such as the *Tractor* and *Street Cart* are specific examples where the IMNet is able to learn more descriptive feature representations that help to distinctly localize and identify objects.

4.4.2 Ablation of Hyperparameters of the SCR

The Self-Correlational attention component has two major hyperparameters [13]. The first hyperparameter *planes* affects the local window size of the self-correlational tensor, while the dimension of the input and output embedding is determined by the *dim* hyperparameter. We vary both these hyperparameters and study the effect these have on the model's performance on novel classes during the few-shot adaptation stage (Table 4).

We observe that increasing the dimension of the self-correlational tensor by increasing the value of *dim* improves performance marginally on base classes but leads to a significant degradation in performance on novel classes. However, varying the value of *planes* to change the size of the local window does not have a substantial effect on base class performance but does affect performance on the novel classes significantly. Since we are trying to maximize performance on the novel classes, we choose a value for the *planes* hyperparameter which allows us to obtain the best performance on novel class data.

4.4.3 Ablation of Hyperparameters in the Cross-Entropy Similarity Loss

The Cross-Entropy Similarity loss is used to increase the inter-class separation between each object. It comprises a single hyperparameter λ which determines the contribution of the $L_{similarity}$ term in the total loss objective. We do not observe any continuous trend in novel class performance for the varying values of λ and settle on a value that maximizes our model's performance on novel classes (Table 4).

Table 4. Ablation for the effect of key hyperparameters planes and dim on novel class performance on IDD-OS. The chosen values for the IMNet approach have been underlined and their associated performance are indicated in **bold**.

Parameter	Value	mAP_{base}	mAP_{base}
planes (dim = 1024)	16	41.1	43.4
	32	**40.8**	**47.8**
	64	40.8	45.1
	128	41.5	41.1
	256	41.1	43.5
dim (planes = 32)	1024	**40.8**	**47.8**
	2048	41.0	41.0
λ (planes = 32, dim = 1024)	0.0	40.1	45.6
	0.25	**40.8**	**47.8**
	0.5	40.2	42.4
	1.0	39.4	43.8

4.5. Resilience against Class Confusion and Catastrophic Forgetting

Figure 3 shows the confusion matrix of class predictions for our proposed approach on all classes present in IDD-

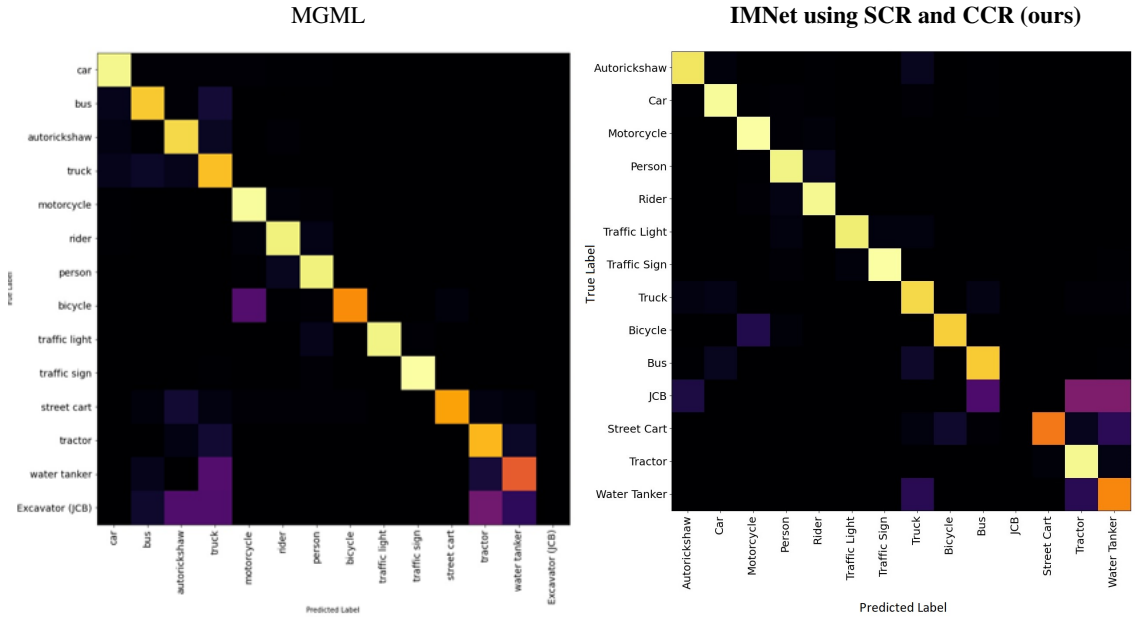

Figure 3. **Confusion Matrix of class predictions**: We show upto 68% reduction in class confusion through our IMNet approach, especially for base classes, over the existing SOTA approach MGML. We also show a reduction in class confusion between base and novel classes at the cost of a slight elevation in class confusion between novel classes.

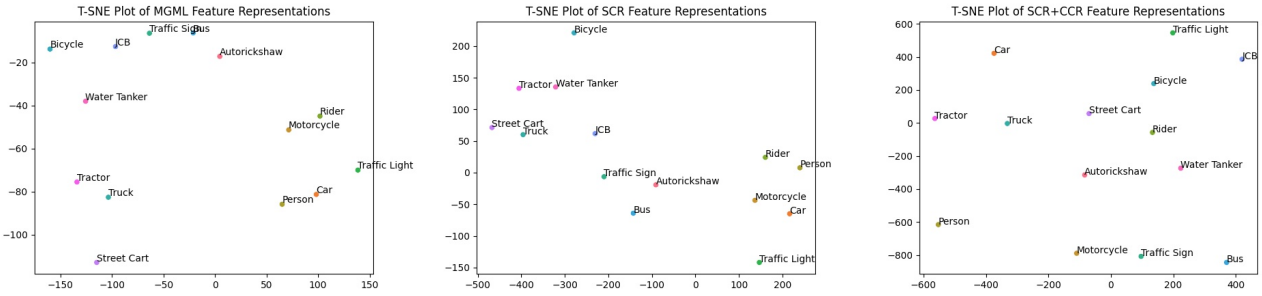

Figure 4. **t-SNE plot of feature representations**: MGML is not able to separate similar looking objects, leading to class confusion. IMNet's SCR and CCR components reduce class confusion by significantly increasing the inter-class separation between the representations.

OS and contrasts it against the existing SOTA approach, MGML. Unlike MGML, where novel classes are confused with one or more base classes, our IMNet shows a significant reduction in confusion between base and novel classes. Overall, our IMNet achieves the least confusion among existing by demonstrating a 68 % reduction in class confusion when compared to MGML. Despite this sharp reduction, object classes like *Excavator* and *street cart* continue to show elevated confusion with other novel classes due to the inter-class bias in IDD. The influence of SCR and CCR modules is viewed spatially through t-SNE plots in figure 4 of the feature representations for each class in the support set, averaged over all shots in the 10-shot setting. We observe that MGML is unable to increase the inter-class separation between visually similar objects, such as *motorcycle* and *rider*. When both the SCR as well as the CCR modules are applied, we see a much larger inter-class separation even between similar-looking objects, thus showing its effectiveness in reducing inter-class variance and inter-class bias.

Table 3 shows the performance of our IMNet architecture on object classes in D_{base}. The combined effect of the SCR and CCR module shows better retention in the performance of the base class when knowledge of novel classes is added to the IMNet network. Overall, our method achieves up to a 3% reduction in catastrophic forgetting over existing SOTA approaches.

5. Conclusion

In this work, we introduced a novel information maximization based few-shot object technique, the Information Maximization Network (IMNet) to overcome catastrophic forgetting in few-shot object detection. Our experiments are conducted on the challenging India Driving Dataset (IDD) and our model slightly outperforms the State-of-the-Art meta learner by 0.2 mAP on base classes and up to 3 mAP on novel classes.

We adapt two attention mechanisms built for few-shot image classification for few-shot object detection which help the model to learn feature representations that encode the most discriminating information about each object and also align these representations to "look" for the most important distinguishing features. We also introduce a new inexpensive loss objective to adjust feature representations and make them more compact and increase inter-class separation, thus reducing class confusion. Our approach is additionally able to improve performance on novel classes without any degradation in base class performance and reduce catastrophic forgetting.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *ICLR*, 2015. 2

[2] Luca Bertinetto, João F. Henriques, Philip H. S. Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. *ICLR*, 2018. 1, 2

[3] Yudong Chen, Chaoyu Guan, Zhikun Wei, Xin Wang, and Wenwu Zhu. MetaDelta: A Meta-Learning System for Few-shot Image Classification. *AAAI*, 2021. 1, 2

[4] Hao Cheng, Yufei Wang, Haoliang Li, Alex C. Kot, and Bihan Wen. Disentangled Feature Representation for Few-shot Image Classification. *CoRR*, abs/2109.12548, 2021. 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. *CVPR*, 2009. 6

[6] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision*, 111, 2015. 2

[7] Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun. Generalized Few-Shot Object Detection without Forgetting. *CVPR*, 2021. 2

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. *ICML*, 2017. 1

[9] Ross B. Girshick. Fast R-CNN. *CVPR*, 2015. 1

[10] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *ICCV*, 2013. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *CVPR*, 2015. 5

[12] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. *CVPR*, 2016. 1

[13] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. Relational Embedding for Few-Shot Classification. *ICCV*, 2021. 2, 3, 4, 6, 7

[14] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, R. Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten Digit Recognition with a Back-Propagation Network. *NeurIPS*, 1990. 1

[15] Hojun Lee, Myunggi Lee, and Nojun Kwak. Few-Shot Object Detection by Attending to Per-Sample-Prototype. *CoRR*, 2021. 2

[16] Aoxue Li and Zhenguo Li. Transformation Invariant Few-Shot Object Detection. *CVPR*, 2021. 2, 4

[17] Zhenguo Li, Fengwei Zhou, Fei Chen, and Hang Li. Meta-SGD: Learning to Learn Quickly for Few Shot Learning. *CoRR*, abs/1707.09835, 2017. 1, 2

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature Pyramid Networks for Object Detection. *CVPR*, 2017. 5

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. *ECCV*, 2014. 2

[20] Jiawei Ma, Hanchen Xie, Guangxing Han, Shih-Fu Chang, Aram Galstyan, and Wael Abd-Almageed. Partner-Assisted Learning for Few-Shot Image Classification. *ICCV*, 2021. 3

[21] Anay Majee, Kshitij Agrawal, and Anbumani Subramanian. Few-Shot Learning for Road Object Detection. In *AAAI Workshop on Meta-Learning and MetaDL Challenge*, 2021. 1, 2, 5

[22] Anay Majee, Anbumani Subramanian, and Kshitij Agrawal. Meta Guided Metric Learner For Overcoming Class Confusion in Few-Shot Road Object Detection. In *NeurIPS Workshop on Machine Learning For Autonomous Driving*, 2021. 1, 5, 6, 7

[23] Luis Perez and Jason Wang. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. *CoRR*, abs/1712.04621, 2017. 1

[24] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal networks. *NeurIPS*, 2015. 1, 5

[25] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental Learning of Object Detectors without Catastrophic Forgetting. *ICCV*, 2017. 1, 2

[26] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition, 2015. 1

[27] Elliott Skomski, Aaron Tuor, Andrew Avila, Lauren A. Phillips, Zachary New, Henry Kvinge, Courtney D. Corley, and Nathan O. Hodas. Prototypical Region Proposal Networks for Few-Shot Localization and classification. *NeurIPS*, 2020. 3

[28] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning. *NeurIPS*, 2017. 1, 2

[29] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding. *CVPR*, 2021. 6

[30] 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. *CVPR*, 2017. 1

[31] Girish Varma, Anbumani Subramanian, Anoop M. Namboodiri, Manmohan Chandraker, and C. V. Jawahar. IDD: A Dataset for Exploring Problems of Autonomous Navigation in unconstrained environments. *WACV*, 2019. 1, 2, 5

[32] Oriol Vinyals, Charles Blundell, Timothy P. Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. *NeurIPS*, 2016. 1, 2

[33] Xin Wang, Thomas E. Huang, Trevor Darrell, Joseph E. Gonzalez, and Fisher Yu. Frustratingly Simple Few-Shot Object Detection. *ICML*, 2020. 5, 6

[34] Yaqing Wang and Quanming Yao. Few-shot Learning: A Survey. *ACM Computing Surveys*, 2019. 1

[35] Nicolai Wojke and Alex Bewley. Deep Cosine Metric Learning for Person Re-Identification. *WACV*, 2018. 1, 2

[36] Aming Wu, Yahong Han, Linchao Zhu, Yi Yang, and Cheng Deng. Universal-Prototype Augmentation for Few-Shot Object Detection. *ICCV*, 2021. 2

[37] Yang Xiao and Renaud Marlet. Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild. *ECCV*, 2020. 4, 5, 6

[38] 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. *ICCV*, 2019. 1, 2, 4, 6

[39] Weilin Zhang and Yu-Xiong Wang. Hallucination Improves Few-Shot Object Detection. *CVPR*, 2021. 2