OpenSceneVLAD: Appearance Invariant, Open Set Scene Classification

William H. B. Smith [1],
Michael Milford [2],
Klaus D. McDonald-Maier [1],
Shoaib Ehsan [1] and R. B. Fisher [3] [*†‡§]

March 1, 2022

Abstract

Scene classification is a well-established area of computer vision research that aims to classify a scene image into pre-defined categories such as playground, beach and airport. Recent work has focused on increasing the variety of pre-defined categories for classification, but so far failed to consider two major challenges: changes in scene appearance due to lighting and open set classification (the ability to classify unknown scene data as not belonging to the trained classes). Our first contribution, SceneVLAD, fuses scene classification and visual place recognition CNNs for appearance invariant scene classification that outperforms state-of-the-art scene classification by a mean F1 score of up to 0.1. Our second contribution, OpenSceneVLAD, extends the first to an open set classification scenario using intra-class splitting to achieve a mean increase in F1 scores of up to 0.06 compared to using state-of-the-art openmax layer. We achieve these results on three scene class datasets extracted from large scale outdoor visual localisation datasets, one of which we collected ourselves.

[*] This work was supported by the UK Engineering and Physical Sciences Research Council through Grants EP/R02572X/1, EP/P017487/1 and in part by the RICE project funded by the National Centre for Nuclear Robotics Flexible Partnership Fund.

[†1] W. H. B. Smith, K. McDonald-Maier and S. Ehsan are with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, United Kingdom. (e-mail: ws8@hw.ac.uk; kdm@essex.ac.uk; sehsan@essex.ac.uk)

[‡2] M. Milford is with the School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, QLD 4000, Australia. (e-mail: michael.milford@qut.edu.au)

[§3] R. B. Fisher is with the School of Informatics, University of Edinburgh, EH8 9AB, United Kingdom. (e-mail: rbf@inf.ed.ac.uk)

1 INTRODUCTION

Scene classification aims to classify a scene image into a pre-defined scene category (such as playground, beach and airport). It has applications for content based image retrieval [1], robot navigation [2] and disaster detection [3].

Changes in visual appearance make scene classification challenging, particularly outdoors where variations in lighting, weather and season are commonplace. However, until now this problem has remained unaddressed [4]. In contrast, recent research into visual place recognition (VPR) has focused entirely on appearance invariance [5] [6] [7]. VPR is traditionally formulated as an image retrieval task [8] [9] [10] that compares a reference database of place images to a set of query place images. Our first contribution takes inspiration from contrastive learning [11] to fuse and partially retrain VPR [5] and scene classification [12] convolutional neural networks for appearance invariant scene classification.

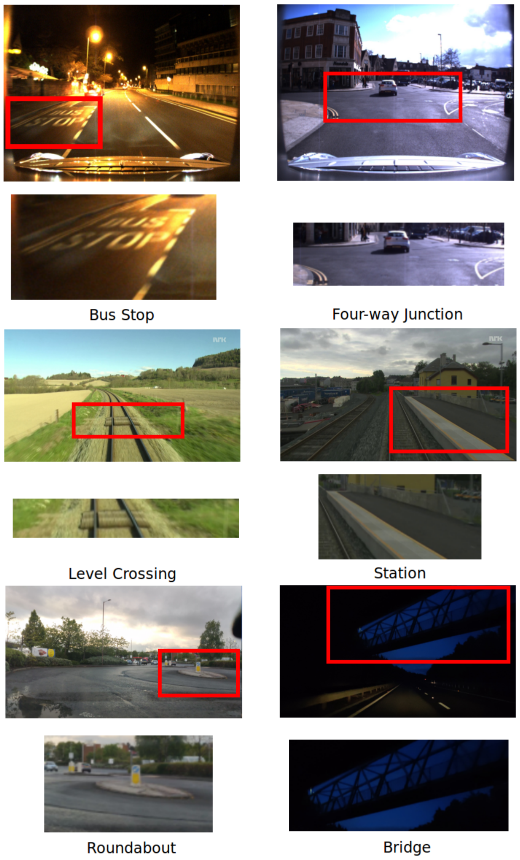
Open set classification (OSC) is generally defined as recognising unknown classes that are not included in the training dataset, but are present at test time. Scene classification is inherently an open set problem because it attempts to describe real world scenes which are unpredictable and vary infinitely. Evidence for this can be seen in the growth of state-of-the-art scene classification datasets to include up to 365 different scene classes [12]. OSC remains a very challenging area of research [13] and therefore current approaches limit themselves to simple computer vision datasets such as CIFAR-10 and MNIST [14] [15]. To our knowledge, this work is the first to consider a significantly more challenging scene classification dataset. Our second contribution uses intra-class splitting [14] to identify atypical training examples of our known scene classes for learning open set scene classification.

For experimentation we identify four different scene classes amongst an open set of unknown scene images in a variety of appearances across three large-scale visual localisation datasets: Oxford RobotCar [16], Nordland [17] and a third Edinburgh dataset we collected ourselves. In summary, our contributions are:

1. A visual localisation dataset we make publicly available covering a 20km traversal of Edinburgh in three different visual conditions with GPS data and labels for 4 scene classes (Section 3.1). We also make available labels for 4 scene classes across three traversals of the Oxford RobotCar and Nordland datasets (`https://github.com/WHBSmith`).

2. A combined scene classification and visual place recognition CNN 'SceneVLAD' trained for appearance invariant scene classification (Section 3.2).

3. An investigation into the significance of open set scene classification as a problem and an extension 'OpenSceneVLAD' to our second contribution using intra-class splitting specifically for this (Section 3.3).



Bus Stop

Four-way Junction

Level Crossing

Station

Roundabout

Bridge

Figure 1: Examples of Oxford (top), Nordland (middle) and Edinburgh (bottom) scene classes, showing the appearance variation within datasets. Class specific areas are outlined in red.

2 BACKGROUND

2.1 Scene Classification

Convolutional neural network (CNN) features used for state-of-the-art scene classification [18] [4] can be grouped into the five categories explored below.

Global features typically use a generic CNN trained on a task such as object detection and then fine tuned for scene classification [12]. These features provide enriched spatial information and global scene properties, but are more vulnerable to background noise [19].

Spatially invariant features usually extract multiple local patches [20] using VLAD or Fisher encoding [21]. These approaches are efficient, but perform poorly when scenes include objects with variable sizes or aspect ratios.

Semantic features use object detection to identify salient contextual regions within a scene [22]. However, the lack of detailed scene labels and the computational requirements of searching these regions [23] limits these approaches.

Multi-layer features combine different resolution features from different CNN layers [24]. For example, DAG-CNN [25] integrates features from different levels of a CNN in a directed acrylic graph. Feature fusion is necessary for this approach and can be done early by extracting multi-layer features and merging them into a comprehensive feature, or late by using supervised learning to ensure features remain sensitive to each target scene class [26].

Multi-view features from multiple complementary CNN models trained on different datasets can be used to create comprehensive scene representations. For example, FOSNet [27] introduces scene coherence loss to fuse object and scene data while Sun et al [28] separately extracts three complimentary representations using object semantics, contextual information and global appearance.

For the base of our SceneVLAD approach we use two global CNN feature networks from Zhou et al's work [12]. '365' is taught to classify 365 scene classes from the Places datasets [12] and '1365' is taught to classify Places and the 1000 ImageNet classes [29]. These CNNs were selected because of their performance on a wide variety of scene classes, the availability of their pre-trained weights and the ease with which their architecture allowed integration of NetVLAD [5] descriptors.

2.2 Visual Place Recognition

VPR descriptors are used in visual localisation and are designed to be invariant to viewpoint [30] [31], environmental [32] [33] and long-term appearance [34] changes. They are largely CNN-based with the most successful using triplet learning [35] [36] to learn a similarity metric for image retrieval and improve relevant feature selection for VPR [37]. VPR descriptors trained specifically for image retrieval are not directly comparable to features extracted from an intermediary layer of a CNN trained on classification, hence our distinction between the use of 'descriptor' and 'feature'.

NetVLAD has been shown to generalise well to a variety of environments and remains the backbone of state-of-the-art VPR descriptors [7]. For this reason we use it to provide the appearance invariance for our network. NetVLAD [5] appends a VLAD layer to a partially frozen VGG16 network pre-trained on the ImageNet dataset and re-trains it for appearance invariant VPR using triplet learning.

2.3 Open Set Classification

Although the task of open set scene classification has yet to be addressed, open set classification (OSC) has been explored more generally. Geng et al's survey [13] of this emerging field distinguishes between a number of different OSC scenarios, the closest to our own is 'open set recognition': the classification of test images as members of the closed or open set. However, we further this in OSC by simultaneously classifying the individual closed classes and the open class. Geng et al broadly categorize deep neural network approaches for OSC as discriminative or generative.

Deep neural network classifiers typically use a softmax cross-entropy loss which is inherently closed set because of its normalised output. Discriminative approaches try to find an empirical threshold to identify low confident outputs as an open class [38]. Bendale and Boult [39] replace the softmax layer in a deep neural network (DNN) with an openmax layer. Once trained normally using cross-entropy, each class is represented as a mean activation vector (MAV) based on correctly classified samples and fitted to a separate Weibull distribution and used to compute a pseudo-activation for each class, plus an open class. Oza and Patel [40] use an encoder to learn closed set classification and a decoder to learn OSC by reconstructing the input, conditioned on class identity.

Generative methods attempt to synthesize and then classify likely examples of an open class. Generative OpenMax [15] uses generative adversarial networks (GANs) to synthesize unknown classes and then explicitly estimate probability over them. Encoder-decoder GANs are also used to create synthetic open set images for training that lie on the opposing side of the true decision boundary, between the known classes and the open set [41]. Recently Schlachter et al [14] suggested intra-class splitting which trains a classifier on the closed set problem. Unlearnt training data and correctly learnt training data classified with a low confidence threshold are relabelled as an open set for re-training the network. We adopt this approach for our task because implementation of GANs for creating scene class data is very challenging [42].

3 METHOD

3.1 Scene Class Labelling

As this is the first specific attempt at appearance invariant scene classification, there are no datasets readily available for use. We therefore source our scene class images from three large outdoor visual localisation datasets. These datasets cover large geographical areas in different appearance conditions (Figure 1) making them perfect for extracting scene classes in a variety of different conditions and providing a challenging open set of unlabelled scene images.

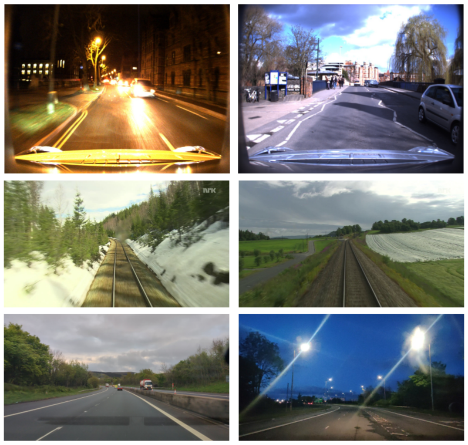


Figure 2: Examples of Oxford (top), Nordland (middle) and Edinburgh (bottom) open set scene classes from the same traversals as Figure 1

Table 1: Dataset, appearance condition traversals used for training and testing with average number of frames per scene class, per traversal.

Dataset	**Traversals**	**Scene Class**	**Frames**
Oxford	Overcast (*train*)	Pedestrian Crossing	170
	Night (*test*)	Bus Stop	129
	Sunny (*test*)	Four-way Junction	152
		T-junction	103
		Open Set	2739
		Total	**3293**
Nordland	Spring (*train*)	Bridge	81
	Winter (*test*)	Level Crossing	87
	Summer (*test*)	Station	292
		Tunnel	186
		Open Set	3970
		Total	**4626**
Edinburgh	Overcast (*train*)	Pedestrian Crossing	53
	Evening (*test*)	Roundabout	135
	Sunny (*test*)	Bus Stop	46
		Bridge	47
		Open Set	843
		Total	**1124**

The first dataset is the 9km urban Oxford RobotCar dataset. The three traversals we use are: 2015-07-03-15-23-28 (overcast), 2014-12-16-18-44-24 (night) and 2015-03-24-13-47-33 (sunny). The second dataset is Nordland, a 763km train journey through rural Norway. The traversals we use are: spring, winter and summer. The final Edinburgh dataset we collected from three traversals of 20km of urban, rural and motorway environments: 20210524 (overcast), 20210526 (evening) and 20210804 (sunny) using a dash-mounted OnePlus 7T recording 4k video at 30fps and a GPS logger app. Each dataset was sampled using a minimum distance between consecutive frames of 0.1 (Oxford), 80 (Nordland) and 2 (Edinburgh) meters to prevent oversampling single scenes, for example when the raw video data stops at a pedestrian crossing.

For each dataset we hand-labelled four scene classes in each traversal based on how frequently and evenly they occur in the dataset and their potential significance for tasks such as visual navigation and collected the remaining images into the open set (Figure 2). We then randomly chose one traversal for training/reference and the other two for testing and fix these for all experiments, as summarised in Table 1. Note, some scene frames represent multiple views of a single scene instance. For example, two scene frames may represent the approach and then traversal of a single roundabout. All possible efforts were made to select the labels reasonably and consistently.

3.2 SceneVLAD: Appearance Invariant Scene Classification

3.2.1 Basic Idea

Successful VPR descriptors are based on a learned similarity metric that represents place images consistently despite changes in their appearance. However, because VPR descriptors are designed for visual localisation each place is represented uniquely, not as belonging to a scene class. Our hypothesis is that, despite this, they can be used to improve appearance invariance for scene classification.

3.2.2 Architecture

To combine two neural networks, one taught for classification and the other for descriptor generation, we take inspiration from recent work on supervised [43] and self-supervised [11] contrastive loss.

Contrastive loss is a distance-based loss function that can be used to partially train a network on a ‘pretext task’ to represent data in a way that would be helpful as input to a later part of the network taught on a ‘downstream task’, such as classification. Usually the pretext task and downstream task would be the same, but in this case we want to combine the knowledge from two different tasks so we use NetVLAD pre-trained for VPR using triplet loss, a variation of contrastive loss, rather than scene classification. A downstream classifier would normally then be appended and trained on the task, but as our pretext and downstream tasks are different we fuse it with the higher levels of an entire scene classification network and then partially re-train the entire fused network for scene classification.

We therefore have two inputs to our network. Firstly, we pass an image to ‘365’ or ‘1365’ networks pre-trained on scene classification and freeze the top 16 layers to produce a 365 or 1365 dimensional output. Secondly, we pass the same image through NetVLAD, but add a small, trainable 1x1 convolutional layer rather than NetVLAD’s native PCA layer to reduce the output dimensions from 4096 to 256 and allow a linear weighting of the most useful features to be learnt, rather than assuming the PCA features are the most useful for our task. Finally, we concatenate the two outputs and pass them through two fully connected layers of size 4096 to a 4-class softmax classification layer (Figure 3). This architecture forces SceneVLAD to learn scene classification with respect to NetVLAD’s appearance invariant image descriptors, thereby improving appearance invariance for scene classification.

We optimized the number of filters used for dimension reduction, the number of final fully connected layers and their width using a grid search. We found that freezing the top 16 layers for re-training 365/1365 produced the best results in all cases. Additionally, we explored using Khosla et al’s supervised contrastive approach [43] to explicitly re-train NetVLAD for our task. However, NetVLAD is only taught for appearance invariance down to the conv5 layer so re-training these layers significantly compromised its performance for the task. We also varied the width of the NetVLAD dimension reduction layer for multiplication of the two networks’ output rather than concatenation, as suggested in [24], but this did not improve performance.

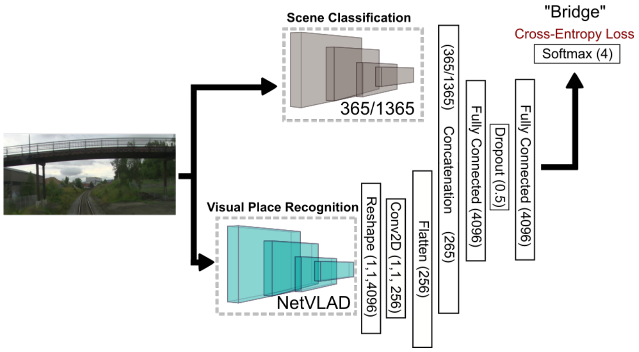


Figure 3: SceneVLAD network architecture fusing of scene classification and visual place recognition for appearance invariant scene classification.

3.3 OpenSceneVLAD: Open Set, Appearance Invariant Scene Classification

3.3.1 Basic Idea

Intra-class splitting [14] was selected to extend SceneVLAD for open set scene classification (OpenSceneVLAD) because we hypothesised it could leverage assumptions about likely open set scene images from training images, such as position of the ground plane and orientation of ambient scenery for improved OSC performance.

3.3.2 Identify Atypical Class Examples

Firstly, SceneVLAD is trained for closed set classification on N number of classes using scene images x_i and the corresponding closed set labels $y_{i,cs}$. The trained network is then used to classify the training images. A user defined proportion of images per class, in this case 30% [14], that cannot be classified correctly, or are classified with the lowest confidence softmax output values are identified as 'atypical' examples. We deviate slightly from the original technique by identifying atypical images proportionally across all training classes. This was necessary because of the small and imbalanced classes.

3.3.3 Generate Open Set Labels

A new set of scalar labels is then generated $y_{i,os}$. For every image, x_i, if deemed atypical $y_{i,os} = N + 1$, otherwise $y_{i,os} = y_{i,cs}$. Relabelling misclassified or unconfidently classified images allows an open set to be generated using only closed set images.

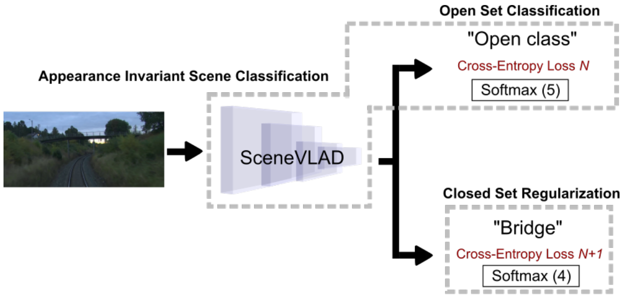


Figure 4: OpenSceneVLAD network architecture demonstrating intra-class splitting on an atypical image.

3.3.4 Create OpenSceneVLAD and Re-train

Two separate softmax layers are used as network outputs (Figure 4), one with N outputs trained for closed set regularization using cross-entropy loss $\mathcal{L}_{cs}$

(Equation 3), the other with $N+1$ outputs trained for OSC also using cross-entropy loss $\mathcal{L}_{os}$ (Equation 2). OpenSceneVLAD is then trained using both losses with the corresponding labels of each image. Closed set regularization helps maintain a high closed-set accuracy by forcing the atypical samples to be correctly classified to their original classes.

3.3.5 Test

At test time the closed set regularization is removed and output from the remaining open set layer is used for open set classification.

3.3.6 Loss Functions

The objective of intra-class splitting is a joint optimization problem (Equation 1) consisting of two individual loss terms for the open set layer and closed set layer

$$\mathcal{L} = \mathcal{L}_{os} + \gamma * \mathcal{L}_{cs}, \tag{1}$$

where $\mathcal{L}_{os}$ is the loss function for the open set layer and $\mathcal{L}_{cs}$ is the loss function for the closed set layer. γ is a hyperparameter to tune the focus on closed set regularization.

Let B be the minibatch size during training. Moreover, $1_{y_i \in y^{(n)}}$ is an indicator function which returns 1 if a given sample x_i with a scalar label y_i belongs to the class $y^{(n)}$ and otherwise returns 0. Based on these notations $\mathcal{L}_{os}$ is a simple $(N+1)$-class categorical cross-entropy loss

$$\mathcal{L}_{os} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{n=1}^{N_{os}} 1_{y_i \in y^{(n)}} \log[P(\hat{y}_i \in y^{(n)})] \tag{2}$$

where $N_{os} = N+1$ and $P(\hat{y}_i \in y^{(n)})$ denotes the predicted probability that sample x_i belongs to the class $y^{(n)}$, i.e. the value of the n-th element of the network's output vector. $\mathcal{L}_{cs}$ is an N-class categorical cross entropy loss

$$\mathcal{L}_{cs} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{n=1}^{N_{cs}} 1_{y_i \in y^{(n)}} \log[P(\hat{y} \in y^{(n)})], \tag{3}$$

where $\mathcal{L}_{cs}$ shares the same notation as $\mathcal{L}_{os}$ and $N_{cs} = N$ is the number of the given known classes.

4 EXPERIMENTS

The experiments in this section comprise an extended ablation study. We begin with two CNNs that make up SceneVLAD and systematically add functionality, demonstrating improvements in appearance invariance and then open set classification.

4.1 SceneVLAD: Appearance Invariant Scene Classification

The purpose of this experiment was to confirm whether our SceneVLAD fusion of NetVLAD VPR descriptors with scene classification networks 365 or 1365 improves appearance invariance for scene classification.

4.1.1 Baselines

For a baseline we train the VPR and scene classification networks (NetVLAD, 365 and 1365) individually to classify the scenes in each of our datasets from Section 3.1. Each network has the same architecture as they do in SceneVLAD (Figure 3) before the concatenation layer, simply with a softmax layer appended.

4.1.2 SceneVLAD

We create two versions of SceneVLAD using the architecture described in Section 3.2.2, each based on scene classification networks 365 and 1365 combined with NetVLAD to create SceneVLAD365 (Sc.VLAD365) and SceneVLAD1365 (Sc.VLAD1365). For SceneVLAD and the baseline networks the top 16 layers of networks 365 and 1365 were frozen for training and NetVLAD was frozen up to its final normalization layer.

4.1.3 Training

All 5 networks were trained to classify the four scene classes from the training traversal of each dataset using the same approach: image augmentation (featurewise normalisation, width and height shifting, horizontal flipping and brightness changes), a minibatch size of 8, a learning rate of 1e-5 and a stratified training/validation split of 80/20. Training was done for 100 epochs with early stopping using a patience of 20 by monitoring validation classification accuracy. For each dataset we repeated the above training five times and averaged the model results.

4.1.4 Testing

We test our classifiers on the scene classes from the remaining two traversals of the same dataset. The actual scenes therefore remain largely constant. This allows any increase in performance to be attributed solely to improved appearance invariance. For evaluation we use the F1 score to compensate for the class imbalance. The change in F1 scores from SceneVLAD and the other base networks designed for scene classification are highlighted in Table 2.

4.1.5 Does SceneVLAD Improve Appearance Invariance for Scene Classification?

The results in Table 2 show a mean increase in F1 scores of up to 0.1 across a variety of scenes when VPR descriptors and scene classification are combined

Table 2: Results of ablation study comparing our SceneVLAD's (Sc.VLAD) fusion of VPR and scene classification with its constituent parts all trained on appearance invariant scene classification. Changes in mean F1 scores (Δ) are Positive and negative.

	Oxford RobotCar				**Nordland**				**Edinburgh**				**Dataset**
	Ped.	Bus	Four	T		Lvl.			Ped.		Bus		**Class**
	Cross	Stop	Junct.	Junct.	Bridge	Cross.	Station	Tunnel	Cross.	Round.	Stop	Bridge	**Mean**
NetVLAD	0.66	0.69	0.67	0.47	0.78	0.81	0.96	0.95	0.04	0.66	0.13	0.05	**0.57**
365	0.65	0.59	0.64	0.49	0.58	0.55	0.93	0.93	0.24	0.84	0.57	0.48	**0.62**
Sc.VLAD365	0.66	0.70	0.65	0.58	0.83	0.87	0.98	0.96	0.51	0.83	0.61	0.47	**0.72**
Δ	**0.01**	**0.11**	**0.01**	**0.09**	**0.25**	**0.32**	**0.05**	**0.03**	**0.27**	**-0.01**	**0.04**	**-0.01**	**0.10**
1365	0.58	0.43	0.64	0.42	0.68	0.68	0.94	0.93	0.37	0.89	0.58	0.65	**0.65**
Sc.VLAD1365	0.67	0.53	0.67	0.66	0.83	0.86	0.97	0.96	0.48	0.89	0.59	0.60	**0.73**
Δ	**0.09**	**0.10**	**0.03**	**0.24**	**0.15**	**0.18**	**0.03**	**0.03**	**0.11**	**0.00**	**0.01**	**-0.05**	**0.08**

in SceneVLAD. This experiment's results demonstrate 365 and 1365, despite being taught in exactly the same way, are outperformed by SceneVLAD's fusion of them with NetVLAD in all but three results. As the main difference between SceneVLAD and the baseline networks is the fusion of scene classification with VPR descriptors this confirms our hypothesis that combining the two does improve appearance invariance for both base scene classification networks.

SceneVLAD specifically outperforms alternatives on scenes most vulnerable to appearance change regardless of underlying class. For example, Oxford's bus stops are typically characterised by only a single bus stop sign and yellow signage on the road, whereas in Edinburgh they often include a large bus shelter (Figure 1). This helps to explain the greater increase in performance for Oxford's bus stop sceneclass with the addition of NetVLAD and further confirm our hypothesis.

4.2 OpenSceneVLAD: Open Set Appearance Invariant Scene Classification

The purpose of this experiment was to examine how scene classification is affected by an open set classification scenario and to examine whether intra-class splitting used in OpenSceneVLAD could improve OSC performance.

4.2.1 Baseline

For this experiment we use the best base scene classification network from the previous experiment (135) and compare it against its SceneVLAD (Sc.VLAD1365) and OpenSceneVLAD (Op.Sc.VLAD1365) derivatives. For further comparison we also add a state-of-the-art openmax layer to 1365 and SceneVLAD (+O) [39] introduced in Section 2.3 for open set classification.

We take 10% of the stratified test images as a validation set for each dataset trial and use these to calculate a confidence threshold optimized to maximize the dataset class mean F1 score. Therefore, networks with outputs equal to the closed set number of classes (1365 and Sc.VLAD1365) predict an open set image if the prediction confidence is below the calculated threshold. Networks with an output that includes the open set (1365+O, ScVLAD1365+O and Op.Sc.VLAD1365) predict an open set image if the open set is predicted, or the prediction is below the calculated threshold. We also consider 1365 and SceneVLAD networks with a threshold of 0 (-base), which are unable to classify any open set images but are included to examine how confidence thresholding affects classification of the closed set classes.

4.2.2 OpenSceneVLAD

We create an OpenSceneVLAD network using 1365 and NetVLAD, per Section 3.3. As before, the top 16 layers of 1365 were frozen for training and NetVLAD was frozen up to its final normalization layer.

4.2.3 Training

We trained all 3 networks (1365, SceneVLAD and OpenSceneVLAD) following the same procedure in our previous experiment (Section 4.1.3), the only change being that OpenSceneVLAD was trained using intra-class splitting per Section 3.3, with 30% of training images being selected as 'atypical'.

4.2.4 Testing

Testing was exactly the same as in Section 4.1.4 with one major difference: the test set now included all the open set images from each dataset's traversal that were unlabelled, as seen in Table 1. Note, we report the individual class F1 scores, but only for completeness as all approaches, other than '-base', are heavily affected by the choice of threshold. For example, '1365+O' is reported with only the open class having a non-zero F1 score, but there are several values of the threshold that generate higher F1 scores for the closed set classes, but result in a lower dataset class mean.

Table 3: Results of ablation study for open set scene classification comparing 1365 and SceneVLAD (Sc.VLAD) with or without openmax layers (+ O) against our OpenSceneVLAD (Op.Sc.VLAD). Best mean F1 score is highlighted in Green.

	Oxford RobotCar					**Nordland**					**Edinburgh**					**Dataset**
	Ped.	Bus	Four	T	Open		Lvl.			Open	Ped.		Bus		Open	**Class**
	Cross	Stop	Junct.	Junct.	Set	Bridge	Cross.	Station	Tunnel	Set	Cross.	Round.	Stop	Bridge	Set	**Mean**
1365 - base	0.15	0.12	0.14	0.23	0.00	0.14	0.05	0.47	0.23	0.00	0.07	0.36	0.20	0.21	0.00	**0.16**
1365	0.10	0.12	0.18	0.32	0.27	0.34	0.07	0.51	0.35	0.00	0.13	0.34	0.12	0.26	0.17	**0.22**
1365 + O.	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.86	**0.18**
Sc.VLAD1365 - base	0.16	0.16	0.23	0.38	0.00	0.21	0.13	0.32	0.43	0.00	0.20	0.66	0.09	0.46	0.00	**0.23**
Sc.VLAD1365	0.16	0.23	0.33	0.37	0.47	0.29	0.09	0.61	0.49	0.40	0.17	0.65	0.07	0.32	0.35	**0.33**
Sc.VLAD1365 + O.	0.19	0.13	0.40	0.38	0.78	0.37	0.14	0.59	0.56	0.70	0.04	0.60	0.06	0.00	0.82	**0.38**
Op.Sc.VLAD1365	0.15	0.15	0.33	0.33	0.82	0.47	0.14	0.67	0.74	0.80	0.27	0.67	0.15	0.28	0.66	**0.44**

4.2.5 How Big a Challenge is Open Set Scene Classification?

The results in Table 3 show a significant decrease in F1 classification score of even known classes compared to Table 2, confirming our hypothesis that the introduction of open set scene images makes this task very challenging due to false positives. OSC has more effect on some classes than others. For example, Nordland level crossings are classified only slightly less accurately in closed set classification than tunnels, however in OSC they are clearly much more challenging, because they are more vulnerable to false positive classifications. Interestingly, SceneVLAD (Sc.VLAD1365) on its own significantly improves mean F1 scores for OSC in comparison to 1365.

4.2.6 Does OpenSceneVLAD Improve Open Set Scene Classification?

Confidence thresholding is shown to improve the mean F1 score for both 1365 (0.06) and SceneVLAD (0.1) with very little effect on classification of known classes compared to the base networks. However, adding the openmax layer, to 1365 (1365+O), results in the open and closed classes becoming indistinguishable and every test image being classified as part of the open set, reducing the mean F1 score by 0.04. Applying openmax to SceneVLAD (Sc.VLAD1365+O) resulted in a large increase in OSC accuracy but was at the expense of the closed classes, in particular for the Edinburgh dataset where its F1 score on the closed classes was the second worst but the change in dataset class mean F1 score was an increase of 0.05.

Using intra-class splitting to train OpenSceneVLAD (Op.Sc.VLAD1365) increases mean scene classification F1 scores by 0.06. This result is particularly significant given that the training data and majority of the underlying network architecture is the same used for SceneVLAD. Specifically OpenSceneVLAD outperforms SceneVLAD with an openmax layer (Sc.VLAD1365+O) on every class of the Nordland dataset, most likely because the scene classes we selected include man-made features amongst a largely rural dataset which intra-class splitting was able to exploit effectively for OSC. For the Edinburgh dataset it also finds the best balance between F1 scores on the closed and open classes.

4.2.7 Additional Classes

Although we use a wide variety of scene classes for evaluation the total class number is strictly constrained by the available data. However, SceneVLAD and OpenSceneVLAD could be applied to more classes. NetVLAD and 365/1365 are based on networks originally taught to classify hundreds of different classes indicating a large remaining model capacity. Intra-class splitting was originally used for 10 classes and we speculate that increasing the classes and therefore the open set data, may improve performance further, but we leave this for future work.

5 CONCLUSIONS

This paper presents a contribution towards appearance invariant and open set scene classification. Fusing VPR descriptors with scene classification in our network SceneVLAD increases average appearance invariance F1 classification scores by up to 0.1. SceneVLAD also improves open set scene classification, which is shown to be a very challenging task for current approaches. Extending SceneVLAD to OpenSceneVLAD with intra-class splitting achieves the best mean increase in open set F1 scores of 0.06, compared to using a state-of-the-art openmax layer.

References

[1] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *International Journal of Computer Vision*, vol. 72, no. 2, pp. 133–157, 2007.

[2] J. Hou, H. Zeng, J. Zhu, J. Hou, J. Chen, and K.-K. Ma, "Deep quadruplet appearance learning for vehicle re-identification," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 8512–8522, 2019.

[3] K. Muhammad, J. Ahmad, and S. W. Baik, "Early fire detection using convolutional neural networks during surveillance for effective disaster management," *Neurocomputing*, vol. 288, pp. 30–42, 2018.

[4] L. Xie, F. Lee, L. Liu, K. Kotani, and Q. Chen, "Scene recognition: A comprehensive survey," *Pattern Recognition*, vol. 102, 2020.

[5] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2018.

[6] M. Lopez-Antequera, R. Gomez-Ojeda, N. Petkov, and J. Gonzalez-Jimenez, "Appearance-Invariant Place Recognition by Discriminatively Training a Convolutional Neural Network," *Pattern Recognition letters*, 2017.

[7] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 141–14 152.

[8] O. Vysotska and C. Stachniss, "Lazy Data Association For Image Sequences Matching Under Substantial Appearance Changes," in *IEEE Robotics and Automation Letters*, 2016.

[9] B. Talbot, S. Garg, and M. Milford, "OpenSeqSLAM2.0: An Open Source Toolbox for Visual Place Recognition under Changing Conditions," in *IEEE Int. Conf. on Intelligent Robots and Systems*, 2018.

[10] F. Lu, B. Chen, X.-D. Zhou, and D. Song, "STA-VPR: Spatio-temporal Alignment for Visual Place Recognition," *IEEE Robotics and Automation Letters*, 2021.

[11] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, “A survey on contrastive self-supervised learning,” *Technologies*, vol. 9, no. 1, p. 2, 2021.

[12] Z. Bolei, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 Million Image Database for Scene Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 11 452–1464., 2017.

[13] C. Geng, S.-j. Huang, and S. Chen, “Recent advances in open set recognition: A survey,” *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[14] Y. L. Patrick Schlachter and B. Yang, “Open-set recognition using intra-class splitting,” in *2019 IEEE European Signal Processing Conference (EUSIPCO)*, September 2019.

[15] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, “Generative openmax for multi-class open set classification,” in *British Machine Vision Conference Proceedings 2017*, 2017.

[16] “1 year, 1000 km: The Oxford RobotCar dataset,” *Int. J. of Robotics Research*, 2017.

[17] N. Sünderhauf and P. Protzel, “Are We There Yet? Challenging SeqSLAM on a 3000 km Journey Across All Four Seasons,” in *ICRA*, 2013.

[18] D. Zeng, M. Liao, M. Tavakolian, Y. Guo, B. Zhou, D. Hu, M. Pietikäinen, and L. Liu, “Deep learning for scene classification: A survey,” *arXiv preprint arXiv:2101.10531*, 2021.

[19] L. Herranz, S. Jiang, and X. Li, “Scene recognition with cnns: objects, scales and dataset bias,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 571–579.

[20] M. Cimpoi, S. Maji, and A. Vedaldi, “Deep filter banks for texture recognition and segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3828–3836.

[21] Y. Li, M. Dixit, and N. Vasconcelos, “Deep scene image classification with the mfafvnet,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5746–5754.

[22] X. Cheng, J. Lu, J. Feng, B. Yuan, and J. Zhou, “Scene recognition with objectness,” *Pattern Recognition*, vol. 74, pp. 474–487, 2018.

[23] Z. Xiong, Y. Yuan, and Q. Wang, “Rgb-d scene recognition via spatial-related multi-modal feature learning,” *IEEE Access*, vol. 7, pp. 106 739–106 747, 2019.

[24] P. Tang, H. Wang, and S. Kwong, "G-ms2f: Googlenet based multi-stage feature fusion of deep cnn for scene recognition," *Neurocomputing*, vol. 225, pp. 188–197, 2017.

[25] S. Yang and D. Ramanan, "Multi-scale recognition with DAG-CNNs," in *Proceedings of the IEEE International Conference on Computer Vision.* IEEE, 2015, pp. 1215–1223.

[26] Y. Dong, S. Gao, K. Tao, J. Liu, and H. Wang, "Performance evaluation of early and late fusion methods for generic semantics indexing," *Pattern Analysis and Applications*, vol. 17, no. 1, pp. 37–50, 2014.

[27] H. Seong, J. Hyun, and E. Kim, "FOSNet: an end-to-end trainable deep neural network for scene recognition," *IEEE Access*, vol. 8, pp. 82 066–82 077, 2020.

[28] N. Sun, W. Li, J. Liu, G. Han, and C. Wu, "Fusing object semantics and deep appearance features for scene recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 6, pp. 1715–1728, 2018.

[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition.* IEEE, 2009, pp. 248–255.

[30] A. Khaliq, S. Ehsan, Z. Chen, M. Milford, and K. McDonald-Maier, "A Holistic Visual Place Recognition Approach Using Lightweight CNNs for Significant ViewPoint and Appearance Changes," *IEEE Transactions on Robotics*, vol. 36, no. 2, 2019.

[31] S. Garg, M. Babu, T. Dharmasiri, S. Hausle, N. Sunderhauf, S. Kumar, T. Drummond, and M. Milford, "Look No Deeper: Recognizing Places from Opposing Viewpoints under Varying Scene Appearance using Single-View Depth Estimation," in *ICRA*, 2019.

[32] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, "Deep Learning Features at Scale for Visual Place Recognition," in *IEEE International Conference on Robotics and Automation.* IEEE, 2017, pp. 3223–3230).

[33] M. Zaffar, S. Ehsan, M. Milford, and K. McDonald-Maier, "CoHOG: A Light-Weight, Compute-Efficient, and Training-Free Visual Place Recognition Technique for Changing Environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1835–1842, 2020.

[34] Z. Wang, J. Li, S. Khademi, and J. van Gemert, "Attention-Aware Age-Agnostic Visual Place Recognition," in *IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

[35] E. Hoffer and N. Ailon, "Deep Metric Learning Using Triplet Network," in *International Workshop on Similarity-Based Pattern Recognition*, 2015, pp. 84–92.

[36] M. Lopez-Antequera, R. Gomez-Ojeda, N. Petkov, and J. Gonzalez-Jimenez, "Appearance-Invariant Place Recognition by Discriminatively Training a Convolutional Neural Network," *Pattern Recognition Letters*, vol. 92, pp. 89–95, 2017.

[37] S. Hausler, A. Jacobson, and M. Milford, "Filter Early, Match Late: Improving Network-Based Visual Place Recognition," in *IEEE International Conference on Intelligent Robots and Systems.* IEEE, 2019, pp. 3268–3275.

[38] D. Miller, N. Suenderhauf, M. Milford, and F. Dayoub, "Class Anchor Clustering: a Distance-based Loss for Training Open Set Classifiers," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV).* IEEE, 2021, pp. 3570–3578.

[39] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563–1572.

[40] P. Oza and V. M. Patel, "C2ae: Class conditioned auto-encoder for open-set recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2307–2316.

[41] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 613–628.

[42] Y. Latif, R. Garg, M. Milford, and I. Reid, "Addressing challenging place recognition tasks using generative adversarial networks," in *2018 IEEE International Conference on Robotics and Automation (ICRA).* IEEE, 2018, pp. 2349–2355.

[43] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, 2020.