Pose Aware Fine Grained Visual Classification using Pose Experts

Kushagra Mahajan
IIIT Delhi
kushagra14055@iiitd.ac.in

Tarasha Khurana
NSIT, Delhi
khurana.tarasha@gmail.com

Ayush Chopra
DTU, Delhi
ayushchopra2k14@dtu.ac.in

Chetan Arora
IIIT Delhi
chetan@iiitd.ac.in

Abstract

We focus on the problem of fine-grained visual classification (FGVC). We posit that unreasonable effectiveness of the state-of-the-art in this area is because of similar object categories present in the ImageNet dataset, which allows such models to be pretrained on a much larger set of samples and learn generic features for those object categories. Such advantage may not be present for other object categories not present in the ImageNet, where one is forced to train with much smaller effective sample set. To work within those constraints, we observe that there is an important and often ignored additional structure present in an FGVC problem: the objects are captured from a small set of viewing angles only, for example, frontal, oblique, top view etc. We notice that subtle differences between various object categories are difficult to pick from an arbitrary angle. However, the same objects become easy to identify from a similar pose. We show in this paper that training separate specialized pose experts, focusing on classification from a single, fixed pose, and then combining them in an ensemble style framework is able to successfully exploit the structure in the problem. We demonstrate the effectiveness of the proposed approach on the benchmark Stanford Cars and FGVC-Aircrafts datasets. We also contribute a new dataset for the problem comprising of 1000 objects of footwear spread across 12 categories, each captured from 4 different poses. The complete source code as well as the dataset will be available post publication at `http://to.be.released.later`.

1. Introduction

Fine grained visual classification aims at distinguishing objects into their subclasses. For instance, dogs are categorized into different breeds of dogs [24], and birds are categorized into different families of birds [2, 27]. However, the fine grained distinction between objects often requires addressing two contradictory issues: 1) distinguish classes which have very subtle differences between them 2) manage the large intra-class variation that arises due to different shapes as well as poses of the target objects. Though, in principle, automated learning of inter and intra class variations are possible with an end to end trained deep neural network, doing it in practice for fine grained classification has been difficult because of lack of large datasets.

Figure 1: To work with limited dataset of generic objects with large intra class and lesser inter class variation in a typical FGVC problem, we propose to exploit fixed viewpoints often present in such problems. (a) and (b) shows images of two clogs from different viewpoints. (c) and (d) show images of a clog and a shoe. Notice the difference in (a) and (b) and similarity of (c) and (d). We observe that the objects become easy to identify when seen from a same viewpoint. (e) and (f) shows the images of two clogs from same viewpoint. (g) and (h) shows images of a clog and shoe again from a same viewpoint. This motivates us to create an ensemble of pose experts in the proposed model.

In our work, we focus on the pose-aware dimension of the fine grained visual classification (FGVC) problem. We observe that in most of the FGVC problems, the number of viewpoints are typically few and fixed, for example, frontal, oblique, top view etc. Further, the subtle differences between various object categories are difficult to pick from an arbitrary angle, but become much simpler when done from a similar pose. For example, consider the problem of classi-

fication for clogs vs casual shoes as shown in Fig. 1. There are large variations between images of a clog when seen from different viewpoints, and on the other hand a clog and a shoe may look very similar from different views. However, the task becomes easier if we exploit the pose structure inherent in the problem and see the objects from same pose.

We exploit the viewpoint structure in the problem and suggest a novel neural network containing a *Pose Expert* specializing for prediction from a specific pose alone. Our hypothesis is that the presence of multiple poses in the data confuses the state-of-the-art deep network models compared to when the same data is segregated pose-wise. The pose specific nature of the expert reduces the intra class variance with respect to inter class variations and helps in training the expert even with a limited dataset. We also train a separate network for detecting the pose of an object. We show that an ensemble of various such pose experts pretrained for a single pose, in conjunction with the pose detection network is able to effectively exploit the fixed viewpoints structure in the problem and give superior performance compared to the state-of-the-art which do not use the cue.

The specific contributions of this work are as follows:

1. We hypothesize that the success of the state-of-the-art FGVC techniques is largely due to generic features learnt on a much larger dataset.

2. We propose to exploit novel pose aware structure for FGVC problems. We show that the proposed model containing an ensemble of pose specializing experts in conjunction with pose detection stream improves the state-of-the-art on the standard benchmarks. The improvement increases with lesser representation of the tested objects in standard datasets used for pretraining, confirming our hypothesis above. Note that, in contrast to the state-of-the-art, we neither align the pose, nor attempt to find parts of the object.

3. To further validate our hypothesis, we contribute a new small dataset containing objects of footwear. We chose the category because of lesser number of samples in the benchmark ImageNet dataset [5]. The dataset contains 1000 annotated images in 12 footwear categories scraped from various online stores. Each object has been captured from 4 different viewpoints: 90° facing left, 45° facing left, 45° facing right and 90° facing right. The improvement of proposed technique over state-of-the-art is greater than on cars and aircrafts datasets.

2. Related Work

Fine grained image classification problems have become popular over the past few years particularly on trees, flowers, leaves, butterflies, fish and dog datasets [1, 4, 13, 16, 18, 20, 23, 24, 28, 36]. Compared to generic object recognition, fine grained recognition benefits more from learning critical parts of the objects that can help align objects of the same class and discriminate between neighbouring classes [6, 8, 37]. Majority of the existing approaches have been focusing on localizing and describing discriminative object parts or by explicitly extracting features at landmark points particularly. The part-based one-vs-one feature system [1] is an example of this, where parts-based features are progressively selected to improve classification. An alternative is the deformable parts-model [9, 11] which obtains a combined feature from a set of pre-defined parts. Zhang *et al*. [38] extract pose-normalised features based on weak semantic annotations to learn cross-component correspondences of various parts. Region proposal methods combined with a DCNN have been shown to more accurately localise object parts [36]. Although the part-based approaches are fully automatic, a lot of manual annotation of the data is required. Various works [7, 31, 32, 30] attempt to find parts of the image that are discriminative, without explicit part labels, but cannot achieve the accuracy of a supervised part-based approach.

Ge *et al*. [10] use an approach similar to ours where they have partitioned the data into K non-overlapping sets of similar images and learnt an expert DCNN for each set. They have reported a relative improvement of 12.7% and achieved state-of-the-art results on two datasets: Caltech-UCSD-2011 (CUB200-2011) [27] and Birdsnap [2]. The work of Lin *et al*. [19] consists of two feature extractor models that obtain local pairwise feature interactions in a translation invariant manner which is particularly useful for fine grained categorization. It gives state-of-the-art 84.1% accuracy on the CUB-200-2011 dataset requiring only category labels and no bounding boxes at train time.

Much of the work on deep learning in pose aware problems has been done on face recognition [22, 25, 33, 34, 40, 43]. The area of pose based fine grained classification on the other popular datasets like Caltech-UCSD-2011 [27] and Stanford Cars dataset [14] of fine grained classification has been relatively little explored. A major motivation for our work comes from Pose Aware Models (PAMs) for face recognition proposed by Masi *et al*. [22], an approach that tackles pose variations by multiple pose-specific models and rendered face images, however unlike ours they specifically fix the pose. Bourdev *et al*. [3] have introduced poselets, whereas Zhang *et al*. [40] use Pose Invariant Person Recognition (PIPER) method, to account for pose variations. Another work similar to ours, is by Zhang *et al*. [39], who have used pose aligned network for inferring human attributes such as gender, hair style, clothes style, expression, and action. In a thesis, similar to ours, they have also proposed that deep convolutional networks trained on large

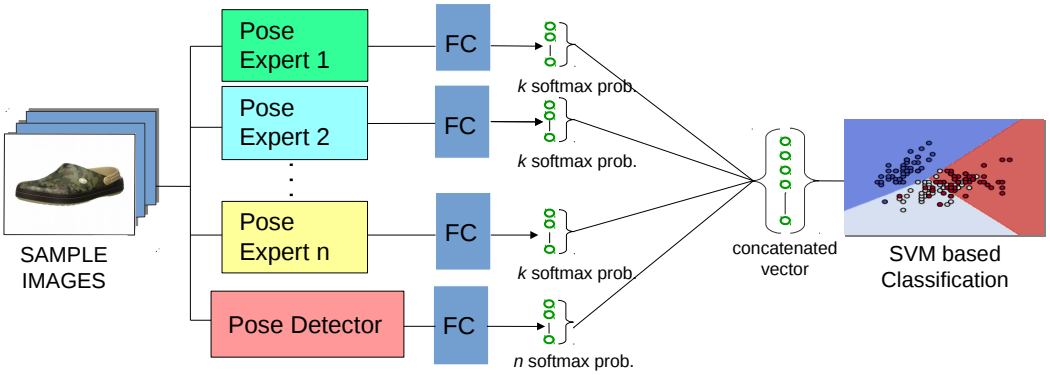

Figure 2: Proposed Network Architecture: We propose an ensemble of pose experts for the targeted fine grained categorization task. Our hypothesis is that it's difficult for a single network to learn inter and intra class variations from small dataset often found in such tasks. On the other hand it's easy to train a network for each pose or viewpoint separately and then combine them to classify from an arbitrary pose. We use standard deep and shallow network architectures for pose experts and pose detector. Our experiment indicates that the proposed model even with shallow pose experts can improve the accuracy from single deep network model.

datasets are generic and can help in other visual recognition problems but may under-perform compared to conventional methods which exploit explicit pose or part-based normalization. However, they use a different strategy than ours and combine a part-based representation with convolutional nets in order to obtain the benefits of both the approaches.

3. Proposed Approach

In this paper, we propose the idea of 'Pose Experts' for pose aware fine grained image classification. We define a Pose Expert as a network trained on only one particular pose-specific data in a fine grained environment which is essentially, distinctive by class and consistent by pose. To aid prediction by the proposed Pose Experts, an additional meta-network is used which is trained for identifying a specific pose of the supplied image. An ensemble of these networks is used to obtain the final prediction. Fig. 2 gives a pictorial description of the proposed network architecture. We give further details of the proposed model below.

3.1. Network Architecture

We have experimented with various shallow (LeNet [17]), deep (AlexNet [15] and VGG16 [26]), and very deep (ResNet-50 and ResNet-101 [12]) CNN architectures. We have also tested with a 'Reduced VGG16' model, obtained by removing fc6 and fc7 layers in the original VGG16 model and 'Reduced Alexnet' which has been arrived at by removing fc7, the last fully connected layer in AlexNet. The ReLU activation function has been used throughout. For all experiments, we have used pre-trained ImageNet dataset [5] weights for AlexNet, VGG16, ResNet models while fine-tuning parameters for all layers. This was the only viable training option given the small datasets available. LeNet5, being a rather shallow network, was trained for all layers from scratch.

Pose Expert: VGG16 has been used as the Pose Expert for the benchmark datasets after experiments with ResNet-50, VGG19, ResNet-101 and other deeper networks. The final layer (fc8) has been retrained and the entire network is finetuned to obtain the best possible accuracy. We took a base learning rate of 0.0001, step decay policy, and a decay of 0.1 along with a regularisation of 0.05. The network has a dropout of 0.5 in fc6 and fc7 layers to eliminate any overfitting. The batch size used was 30 during training and 20 during the testing phase.

The most efficient Pose Expert in the case of the smaller dataset (Footwear dataset in our case) was observed to be R-VGG16 network. A smaller network (i.e. R-VGG16) gives better accuracy for this dataset as the number of classes to be modelled is fewer and deeper networks tend to overfit. The hyperparameters are similar as above and only the regularisation parameter was reduced to 0.001. The batch sizes are 30 and 20 during training and testing respectively.

Pose Detector: We tested with various standard architectures from the 8-layer Alexnet to 101-layer ResNet model. VGG16 gave the best results with parameter values similar to those for the pose expert networks. Fc8 with nodes equal to the number of poses was retrained and the entire network was finetuned.

3.2. Feature Concatenation

For fluent explanation in this section, it is assumed that the number of poses under examination are n and the number of fine grained categories are k.

Figure 3: Representative images from FGVC-Aircrafts dataset [21] which was divided into two poses: left facing, right facing.

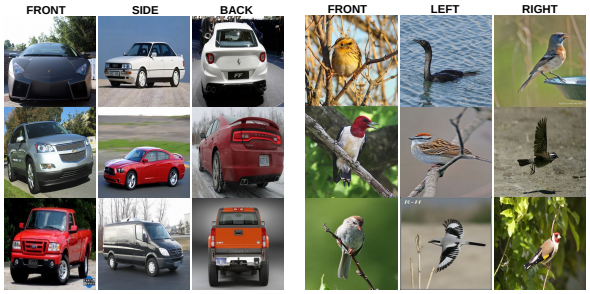

Figure 4: (Left) Representative images from the Stanford Cars dataset [14] which was divided into three poses: front, side, back. (Right) Images from the CUB200-2011 birds dataset [27] divided into three poses: front, left facing, right facing.

For classifying an image from an arbitrary pose we create an ensemble of pose experts. For this purpose, we input the test image from an arbitrary pose (not necessarily the view for which the expert is trained for) to each of the n experts. We then concatenate the k-dimensional vectors containing class-wise confidence scores into a single $n * k$ dimensional feature vector. We also train a separate classifier for identifying the pose, called a 'Pose Detector'. n-dimensional score vector from this meta-classifier is then concatenated with an $n*k$ dimensional vector from the pose experts. We then learn an SVM based classifier from this $n * k + n$ dimensional vector for k-output classes.

In our experiments, we have also tried using the outputs of different fully connected layers from different network models, followed by dimensionality reduction using PCA. However, the proposed configuration yielded the best results. We show further in the experiments section that the proposed network formed by the combination of pose experts, together with a meta-network aiding pose identification, is able to achieve higher accuracy than a single network trained for all the poses together.

4. Dataset and Evaluation

We have compared our approach with various state-of-the-art techniques on different benchmark datasets apart from creating a new dataset containing 1000 images of footwear. The dataset has helped us bring out certain key contributions of the proposed approach. We describe below the datasets and evaluation methodology used for comparison.

FGVC-Aircraft dataset: The Aircrafts dataset consists of 10,000 images spanning 100 models and was introduced as a part of the FGComp 2013 challenge [21]. All models are visually distinguishable, even though in many cases the differences are subtle, making classification challenging. Unlike the other FGVC datasets such as CUB birds dataset [27], airplanes occupy a much larger portion of the image and are rigid. We are thus able to define a definite pose for each of the images. Images were divided into 2 poses: left facing and right facing as shown in Fig. 3. This choice of poses was used because most of the images in the dataset are either facing entirely left or entirely right. Any other choice of pose would have led to insufficient number of images in the pose categories for appropriate training.

One can argue that even if there was only the image of an aircraft available in a sideway (left or right) looking manner, the other image could have been generated by flipping and augmenting such training data. However, our argument is rather different, where we claim that it is still better to train separate networks for left and right facing views. Our claim is that training a single network for both views is a sub-optimal choice when the number of samples are few and inter class variance is low.

Stanford Cars dataset: We also validate our approach on the Stanford Cars dataset [14], which contains 16,185 images of 196 car categories. In our experiments we adopt the predefined 50-50 split for training and testing with 8,144 images for training and the rest for testing. Images from this dataset were divided into 3 poses: front facing, side facing and back facing as shown in Fig. 4.

Each Pose Expert receives pose-specific data and while training a single network for the entire dataset, all poses are supplied together. We maintain equal proportion of various poses in train and test splits.

CUB-200-2011 dataset: CUB-200-2011 is a 200 bird species recognition dataset which contains 11,788 images. It is considered as one of the most competitive fine grained visual classification datasets due to relative scarcity of discriminative features between 200 categories of birds species. Intra class variance is high due to variation in lighting and background and extreme variation in pose (e.g., flying birds, swimming birds, and perched birds that are partially occluded by branches). We use the default training/test split, which gives us around 30 training examples

per class and the dataset is manually segregated into 3 poses: front, left facing and right facing. Some examples have been shown in Fig. 4. Each image is annotated with bounding box, part location, and attribute labels. However, in the proposed scheme we use only the class labels.

Footwear Dataset: We initiated our research using the popular UT-Zappos footwear dataset [35] which has about 50,000 images and provides a significantly large benchmark for analysis. However, the lack of diversity with respect to the poses in the dataset rendered it unfit for use in our research. Consequently, around 1000 images were scraped from online stores such as *amazon* corresponding to 12 classes for four different poses. The classes spanned across: Ankle Boots, Knee High Boots, Formal Shoes, Casual Shoes, Sandals, Slippers, Ballerinas, Boat Shoes, Clogs, Ethnic Chappal, Ethnic Juti, Heels. The four poses used were: Facing Left, Facing Right, Diagonal Facing Left and Diagonal Facing Right. The pose-specific data is mutually exclusive. The images have a plain white background with no occlusion present from any other object and the footwear under consideration occupies a majority portion of the image. Fig. 5 shows some representative images from the contributed dataset.

5. Experiments and Results

The proposed method was implemented in Python and C++ using Caffe library. The experiments were carried out on PCs with mid-level graphics cards; NVIDIA Quadro P5000 and NVIDIA TITAN X and a multi-core 2.1 GHz CPU.

5.1. Implementation Details

Pre-defined architectures were used for experimental testing except for the 'Reduced VGG16' and 'Reduced Alexnet' models. We have used the LeNet5 architecture and subsequently, refer to it as LeNet throughout the paper. For the benchmark datasets, we use two protocols for evaluation: one where the object level bounding box is not provided either at training or at test time ie. *'\bbox'* and the other *'bbox'* where the object level bounding box is used both during the training and testing phases.

To generate sufficient amount of data in order to train our networks efficiently, we augmented the data by techniques like resizing, adding salt and pepper noise, and blurring. Since our problem involved pose monitoring, we had to be careful not to apply the most common augmentation strategies like flipping and rotation for pose experts that could alter the inherent pose-based nature of the problem. However, we have used flipping and rotation while augmenting data for training the compared techniques.

In the following subsections, we will elucidate two novel concepts that we have validated on multiple benchmark datasets and will compare our pose-aware approach against the state-of-the-art methods to establish its viability while analysing the extent of representation of each of the datasets in ImageNet.

Classes	**Single Network**			**PE Network**		
	LeNet	AlexNet	VGG16	LeNet	AlexNet	VGG16
4	72.2	87.3	88.1	80.7	90.5	90.8
8	63.7	74.2	73.2	71.3	82.3	82.7
12	52.1	73.4	72.1	59.6	79.1	79.3

Table 1: Performance of Pose Experts based architectures in comparison to single network for all the poses. We have used 150 images per pose for classification into 4/8/12 classes. The proposed model improves the performance irrespective of the underlying architecture used. Here 'PE Network' denotes Pose Ensemble Network.

Classes	**Single Network**		**PE Network**	
	R-AlexNet	R-VGG16	R-AlexNet	R-VGG16
4	88.3	88.8	93.1	94.1
8	77.5	78.6	84.5	86.3
12	76.2	77.5	82.6	83.5

Table 2: We analyze if an ensemble of pose experts is useful with reduced capacity networks, which can potentially be trained with small samples available in an FGVC dataset. We experiment with smaller networks: reduced AlexNet (R-Alexnet) and reduced VGG16 (R-VGG16) arrived after pruning last few FC layers. The table shows that pose experts are useful even with smaller networks.

5.2. Analysis and Characterization

Pose Experts Vs Single Network: One of the main hypothesis of the current work is to establish the effectiveness of training and merging multiple pose experts to outperform a single network on a dataset that contains distinctly segregable images based on their pose. A single network is supplied with all pose-related data together during training while the pose experts specialize on specific poses. These pose-wise predictions from multiple pose experts are then combined to obtain the final prediction value which is substantially greater than the accuracy of a single network.

We validate this concept first on the Footwear dataset. In the first set of experiments, we trained single networks that contained images from all poses and all classes in equal proportions. 600 distinct images were used in all, 150 for each pose. For 12 footwear classes and a combination of all poses, LeNet obtained an accuracy of 52.1%, AlexNet 73.4% and VGG16 72.1%. Following this, we independently trained pose experts for each of the four poses. Each

Figure 5: Representative images from the contributed footwear dataset. Please refer to the text for details of the dataset.

Poses	Single Network			PE Network		
	L	A	R-VGG16	L	A	R-VGG16
2	54.8	75.3	79.1	60.7	79.8	83.9
4	52.1	73.4	77.5	59.6	79.1	83.5

Table 3: Comparison of Pose Experts and Single Network for different number of poses and 12 classes. The performance of proposed model remains almost same with increase in number of poses, while the advantage with respect to single network increases with increase in number of poses. Here 'L' stands for LeNet, 'A' stands for AlexNet and 'PE Network' stands for Pose Ensemble Network.

	birds		cars		aircrafts	
	\bbox	bbox	\bbox	bbox	\bbox	bbox
Pose 1	75.5	76.8	89.3	93.8	83.2	85.1
Pose 2	77.1	77.9	88.5	92.6	82.7	84.3
Pose 3	78.2	79.1	84.3	87.5	-	-
Pose Detector	93.4	95.3	96.9	97.6	98.1	98.5
Pose Ensemble	76.3	78.4	87.9	92.0	82.5	83.9
Single Net	70.4	76.4	79.8	-	74.1	-

Table 4: Performance of individual Pose Experts and Pose Detector on the benchmark datasets. '\bbox' denotes experiments without bounding box annotation. The proposed ensemble of pose detectors and experts improves the single network performance on all the benchmark datasets. For birds, pose 1, 2 and 3 are front, left, right, for cars these are front, side, back and for aircrafts these are left and right respectively.

pose expert received 150 distinct images each. For a particular test, all pose experts were constituted by the same architecture. Table 1 summarizes the comparison between the pose accuracy of the pose experts based ensemble and that of the single network.

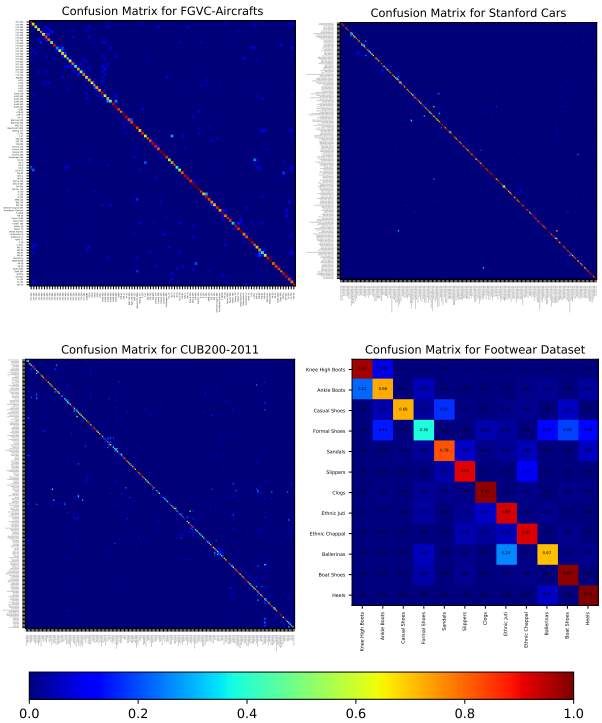

Figure 6: Normalised confusion matrices for FGVC-Aircrafts, Stanford Cars, CUB200-2011 and our Footwear datasets. Please zoom in the pdf to read the fine text in the matrices.

Shallow Pose Experts: Having highlighted the usefulness of pose experts over a single network, the proposed work also indicates the viability of replacing a state-of-the-art single deep network with multiple smaller pose experts. This is potentially useful in fine grained classification tasks where the datasets are often very small with lower number of classes. An ensemble of shallower networks with less number of trainable parameters is thus able to outperform the single deeper networks. We have used reduced AlexNet and reduced VGG16 as representatives of shallow networks, in which the last fully-connected layers have been removed. Since most of the parameters lie in the fully-connected lay-

	birds		cars		aircrafts	
	\bbox	bbox	\bbox	bbox	\bbox	bbox
Proposed	76.3	78.4	87.9	92.0	82.5	83.9
MixDCNN [10]	-	81.1	-	-	-	-
BCNN [19] \ft	80.1	81.3	83.9	-	78.4	-
BCNN [19] ft	84.1	85.1	91.3	-	84.1	-
BGL [42]	75.9	80.4	86.0	90.5	-	-
SCDA [29]	80.5	-	85.9	-	79.5	-

Table 5: Performance comparison with state-of-the-art on standard datasets. '\ft' denotes without finetuning and '\bbox' denotes without bounding box annotation.

ers, this leads to significant decrease of trainable parameters. From Table 1 and 2, we note that a single AlexNet produced an accuracy of 73.4% whereas a set of reduced VGG16 pose experts outperformed the same significantly by 10.1%.

Effect of Number of Poses: We also tested the performance of our system for a different number of poses in the footwear dataset. These results have been summarised in Table 3.

It may be noted that an alternate model to using the ensemble is to use pose detector followed by appropriate pose expert sequentially. However, unlike the alternative, the proposed model is trainable end to end and gave better performance in our experiments.

5.3. Comparison with State-of-the-Art

For comparisons in this section, we use reduced VGG16 network for the pose experts in the footwear dataset and VGG16 network for the same in benchmark datasets. A consolidated results table for comparison with the state-of-the-art techniques on all the datasets under consideration is provided in Table 5. Common mistakes made by our network are illustrated in Fig. 7. The results on each of the datasets have been analyzed below and their respective confusion matrices are shown in Fig. 6.

Footwear Dataset On the Footwear dataset, Bilinear CNN (DD) without finetune, yields a best accuracy of 78.64% on 12 classes with 4 poses. On finetuning, this figure goes to 81.1%, which is about 2.5% lower than our best result of 83.5% using R-VGG16 as highlighted in Table 2. All experiments on our dataset have been carried out with images of size 224*224 while Bilinear CNN operated on images of twice the resolution i.e. 448*448. When we adopt a resolution of 448*448 in our approach, we get a further increase in accuracy of about $\sim 0.7\%$.

Figure 7: Top two pairs of classes that are most confused with each other from each of the 4 datasets, one dataset per row. Each row contains sample images from the test set which are most commonly confused with the class of the neighbouring column.

FGVC-Aircrafts Dataset Experiments conducted on this dataset showed the maximum improvement from the current state-of-the-art. This can be attributed to the very minimal representation of aircrafts in the ImageNet dataset which has been used for finetuning models in state-of-the-art as well as our approach. From Table 4, it can be seen that a single VGG16 network gives the best accuracy of 74.1%. Our approach outperforms the single network by nearly 8.5%. The Bilinear CNN [19] provides an improved 2 stream VGG16 network which gives an accuracy of 84.1% on the dataset. At the same time, our pose segregation based 2 stream model gives only slightly less accuracy than the bilinear finetuned model proposed in the paper. Our model is able to perform better than the SCDA approach [29] while BGL [42] does not give results on the dataset. When bounding boxes are used, the result from our approach improves to 83.9% from 82.5% obtained without bounding box annotation.

Stanford Cars Dataset Trends obtained for the cars dataset are similar to those obtained in the case of aircrafts. Our approach again performs well on this dataset due to the pose structure in the data. We outdo the single VGG16 network (best performing single network) using our approach by nearly 10%. We again obtain competitive results to Bilinear CNN [19] though we do not outperform their finetuned model proposed in the paper. BGL [42] as well as SCDA [29] are both outperformed by a margin of about 2% each. Results with bounding box are better by around 4% (at 92.0%) indicating that cars have more background clutter compared to any other dataset.

CUB200-2011 Dataset On the CUB200-2011 birds dataset, we fall slightly short of the state-of-the-art accuracy as given by [19]. The primary reason for this seems

to be the lack of consistent poses in the birds dataset. The dataset contains images of birds with extreme variation in pose (e.g., flying birds, swimming birds) and the angle from which the images have been clicked. This makes it highly difficult to narrow down to a fixed number of poses with limited variability. Our pose detector stream also does not give a competent accuracy as in a number of images, the head and torso of many birds are flexibly turned into opposite directions which is a confusing singularity for the network. However, the proposed approach gives similar accuracy as the other state-of-the-art approaches: MixDCNN [10], BGL [42] and SCDA [29]. Fig. 9 shows the average precision-recall curves across the 4 datasets.

For all the benchmark datasets, individual expert accuracies and pose detector performance, their ensemble comparison with the corresponding single network using VGG16 have been mentioned in Table 4.

5.4. Visualization

We use activation maps to highlight the quality of features being learnt by our networks and help in visualising how well the pose experts are able to localize the discriminative image regions which could vary in different poses of the same object. Fig. 8 shows the sample class activation maps for the 4 datasets - footwear, birds, aircrafts, cars respectively - and how the discriminative regions change with the viewpoints. In the first two images of the second row, two different poses of birds of the class 'Blue Jay' from CUB200-2011 dataset focus on different features; beak, feet and tail in the first image whereas wings in the second image. Similarly, in the next two images containing 'Chevrolet Traverse SUV 2012' from Stanford Cars dataset, the pose experts seem to focus on the front and hind wheels, backlight and roof in the first image whereas the headlight and logo in the second image. The images have been generated using technique suggested by Zhou *et al*. [41].

6. Discussion

It seems to us that the success of most state-of-the-art approaches on fine grained visual classification task is really due to the generic features the models are able to learn when trained on a much larger dataset i.e. ImageNet. This allows the network to be able to finetune with even the few samples in the fine grained dataset. However, this advantage is circumspect and may not be necessarily present in all fine grained object categories.

On evaluation, we find that almost 58 classes in ImageNet represent birds, 13 represent cars, 3 represent footwear and 3 represent aircrafts. From our experiments it seems that the proposed approach improved the state-of-the-art with a greater margin as the dataset becomes more and more underrepresented in the ImageNet. Lack of pretrained features forces the models to learn new features

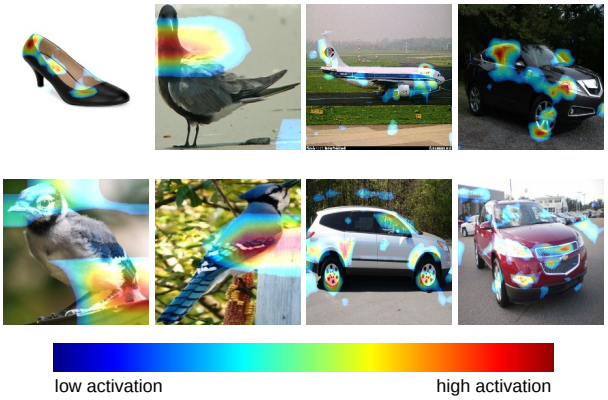

Figure 8: Class Activation Maps. First row shows the activation maps from each of the 4 datasets. Second row shows 2 pairs of images from the same class but different viewpoints and the variation in their discriminative regions. This discriminative information in different poses is directly used by our Pose Experts.

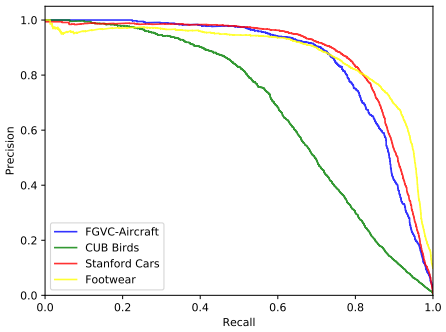

Figure 9: Comparison of the precision-recall curves for the 4 datasets.

from the relatively small number of fine grained samples. The exploitation of structure in the data, such as pose, therefore becomes very important. We speculate that the proposed technique therefore, will become much more important as we have newer dataset for FGVC with objects not represented in larger benchmark datasets.

7. Conclusion

In this paper, we have proposed the idea of using 'Pose Experts' for fine grained visual categorization and shown improvement on the benchmark FGVC datasets. Like all FGVC problems, the annotated dataset available for the problem is small. We posit that it's harder for a single network, deep or shallow, to overcome large intra class variance and small inter class variance as observed from an arbitrary view. This effect becomes particularly profound when the class has limited representation in ImageNet. However, the problem is largely mitigated when viewed from a similar pose. We exploit the observation and train an ensemble of pose experts with an expert for each view, leading to

improvement in accuracy as observed in our experiments. Further, our experiments also show that it is possible to use even shallow network architectures for pose experts, requiring even smaller datasets. We demonstrated this using the contributed footwear dataset. Our results on various benchmark datasets show the effectiveness of our approach when the data contains a definite pose arrangement.

References

[1] T. Berg and P. N. Belhumeur. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 955–962, 2013.

[2] T. Berg, J. Liu, S. Woo Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2011–2018, 2014.

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1365–1372. IEEE, 2009.

[4] Y. Chai, V. Lempitsky, and A. Zisserman. Symbiotic segmentation and part localization for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 321–328, 2013.

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

[6] J. Deng, J. Krause, and L. Fei-Fei. Fine-grained crowdsourcing for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2013.

[7] K. Duan, D. Parikh, D. Crandall, and K. Grauman. Discovering localized attributes for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3474–3481. IEEE, 2012.

[8] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis. Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 161–168. IEEE, 2011.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[10] Z. Ge, A. Bewley, C. McCool, P. Corke, B. Upcroft, and C. Sanderson. Fine-grained classification via mixture of deep convolutional neural networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 1–6. IEEE, 2016.

[11] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 437–446, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[13] J. Jaeger12, M. Simon, J. Denzler, V. Wolff, K. Fricke-Neuderth, and C. Kruschel. Croatian fish dataset: Fine-grained classification of fish species in their natural habitat.

[14] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[16] N. Kumar, P. Belhumeur, A. Biswas, D. Jacobs, W. Kress, I. Lopez, and J. Soares. Leafsnap: A computer vision system for automatic plant species identification. *European Conference on Computer Vision*, pages 502–516, 2012.

[17] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] C. Li, D. Parikh, and T. Chen. Automatic discovery of groups of objects for scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2735–2742. IEEE, 2012.

[19] T.-Y. Lin, A. RoyChowdhury, and S. Maji. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1449–1457, 2015.

[20] J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur. Dog breed classification using part localization. *European Conference on Computer Vision*, pages 172–185, 2012.

[21] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

[22] I. Masi, S. Rawls, G. Medioni, and P. Natarajan. Pose-aware face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4838–4846, 2016.

[23] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[24] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar. Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE, 2012.

[25] A. Sharma, M. Al Haj, J. Choi, L. S. Davis, and D. W. Jacobs. Robust pose invariant face recognition using coupled latent space discriminant analysis. *Computer Vision and Image Understanding*, 116(11):1095–1110, 2012.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[27] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[28] J. Wang, K. Markert, and M. Everingham. Learning models for object recognition from natural language descriptions. In *British Machine Vision Conference*, volume 1, page 2, 2009.

[29] X.-S. Wei, J.-H. Luo, J. Wu, and Z.-H. Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 26(6):2868–2881, 2017.

[30] S. Yang, L. Bo, J. Wang, and L. G. Shapiro. Unsupervised template learning for fine-grained object recognition. In *Advances in neural information processing systems*, pages 3122–3130, 2012.

[31] B. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3466–3473. IEEE, 2012.

[32] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1577–1584. IEEE, 2011.

[33] D. Yi, Z. Lei, and S. Z. Li. Towards pose robust face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3539–3545, 2013.

[34] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim. Rotating your face using multi-task deep neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 676–684, 2015.

[35] A. Yu and K. Grauman. Fine-Grained Visual Comparisons with Local Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2014.

[36] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based R-CNNs for fine-grained category detection. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014.

[37] N. Zhang, R. Farrell, and T. Darrell. Pose pooling kernels for sub-category recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3665–3672. IEEE, 2012.

[38] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 729–736, 2013.

[39] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1637–1644, 2014.

[40] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4804–4813, 2015.

[41] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[42] F. Zhou and Y. Lin. Fine-grained image classification by exploring bipartite-graph labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1124–1133, 2016.

[43] Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-View Perceptron: a Deep Model for Learning Face Identity and View Representations. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 217–225. Curran Associates, Inc., 2014.