POSE AWARE FINE-GRAINED VISUAL CLASSIFICATION USING POSE EXPERTS

Kushagra Mahajan[⋆,1] *Tarasha Khurana*[⋆,1] *Ayush Chopra*[⋆,1] *Isha Gupta*[1] *Chetan Arora*[1] *Atul Rai*[2]

[1]IIIT Delhi [2]Staqu Technologies

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. We observe an important and often ignored additional structure present in an FGVC problem: the objects are captured from a small set of viewing angles only. We notice that subtle differences between object categories are difficult to pick from an arbitrary angle but easier to identify from a similar pose. We show in this paper that training specialized pose experts, focusing on classification from a single, fixed pose, and combining them in an ensemble style framework successfully exploits the structure in the problem. We demonstrate the effectiveness of the proposed approach on the benchmark Stanford Cars, FGVC-Aircrafts, and DeepFashion datasets. To highlight the contribution when the target category features may not be available in a pretrained network, we test on footwear class. We contribute a new 1000 object, 12 category footwear dataset, each object captured from 4 different poses and show significant improvement on this dataset.

***Index Terms*—** Fine Grained Visual Classification, CNN Ensemble, Pose Experts

1. INTRODUCTION

Fine-grained visual classification aims at distinguishing objects into their subclasses. For instance, dogs are categorized into different breeds of dogs [1], and birds are categorized into different families of birds [2, 3]. However, fine-grained distinction between objects often requires addressing two contradictory issues: 1) distinguish classes having 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 is possible with an end-to-end deep neural network, doing it in practice for fine-grained classification has been difficult because of lack of large datasets.

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 classification for clogs vs casual shoes as shown in Fig. 1. There are large variations between images of a clog when seen from different viewpoints, while 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.

Fig. 1: (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 images of two clogs from same viewpoint. (g) and (h) shows images of a clog and shoe from a same viewpoint. This motivates us to create an ensemble of pose experts each specialized to differentiate between the categories from a specific pose.

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 the novel pose aware structure for FGVC problems. We show that the proposed model containing an ensemble of pose specializing experts, in conjunction with the pose detection stream improves the state-of-the-art on the standard benchmarks. As the representation of a dataset reduces in ImageNet, the effect of exploiting the pose related cues becomes more profound, 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

⋆ The first three authors have contributed equally.

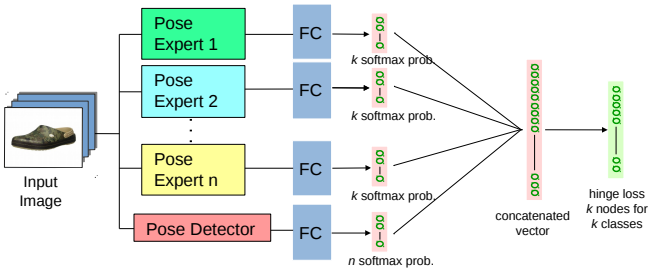

Fig. 2: Proposed Network Architecture.

choose the category because of lesser number of samples in the benchmark ImageNet dataset [4]. The contributed dataset contains 1000 annotated images of 12 footwear categories scraped from various online stores. Each object has been captured from 4 different viewpoints: 90° left, 45° left, 45° right and 90° right. The improvement of the proposed technique over state-of-the-art is greater on this dataset compared to the 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 and dog datasets [1, 5–7]. 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 [7–13]. However unlike our approach, a lot of manual annotation of data is required. Ge *et al*. [14] 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 achieve state-of-the-art results on two datasets: Caltech-UCSD-2011 (CUB200-2011) [3] and Birdsnap [2]. However, they segregate images into subsets based on arbitrary feature distinctions while we exploit the pose variations exclusively. The work of Lin *et al*. [15] 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 training time. A major motivation for our work comes from Pose Aware Models (PAMs) for face recognition proposed by Masi *et al*. [16], an approach that tackles pose variations by multiple pose-specific models and rendered face images, however unlike ours they specifically fix the pose.

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 pose-specific data from only one particular pose 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 as shown in Fig. 2. Note that, an alternative to the proposed architecture is to use the pose detector followed by the appropriate pose expert sequentially. However, unlike the alternative, the proposed model is trainable end-to-end and gave better performance in our experiments. We give further details of our proposed model below[1].

Network Architecture We have experimented with various shallow (LeNet [17]), deep (AlexNet [18] and VGG16 [19]), and very deep (ResNet-50 and ResNet-101 [20]) CNN architectures. Tests were also carried out with a 'Reduced VGG16' model, obtained by removing the fc6 and fc7 layers in the original VGG16 model and 'Reduced Alexnet' which was arrived at by removing fc7, the last fully connected layer in AlexNet. For all experiments, we used pretrained ImageNet dataset [4] weights for AlexNet, VGG and ResNet models while finetuning parameters for all layers. LeNet5 was trained from scratch. Various architectures were tested for Pose Experts and Pose Detector networks. VGG16 gave the best result for the Pose Detector branch throughout. Further details and results are provided in the subsequent sections.

Feature Concatenation Let the number of poses under examination be n and the number of fine-grained categories be k. For classifying an image from an arbitrary pose, we create an ensemble of n pose experts. For this purpose, a test image is sent as input (not necessarily of the view for which the expert is trained for) to each of the n experts. The k-dimensional vectors containing class-wise confidence scores are concatenated into a single $n * k$ dimensional feature vector. n-dimensional score vector from the 'Pose Detector' is then concatenated to form a $n * k + n$ dimensional vector.

Footwear Dataset We initiated our research using the popular UT-Zappos footwear dataset [21] 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 of footwear 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

[1]The complete source code as well as the dataset is available at `http://to.be.released.post.acceptance`.

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 v/s single network for all poses. '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: Usefulness of pose experts with reduced (R) networks.

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. We would like to emphasize here that, the dataset size is kept small deliberately, in consonance with the practical requirements of an FGVC problem where the data is typically scarce and hard to collect. Further, the footwear category is chosen to highlight the effect of under represented classes in the ImageNet dataset. Sample images from our dataset are given in the supplementary material.

4. EXPERIMENTS AND RESULTS

Pre-defined architectures were used for experimental testing except for the 'Reduced VGG16' and 'Reduced Alexnet' models. For the benchmark datasets, we use two protocols for evaluation: one where the object-level bounding box is not provided either at training or testing time i.e. *'\bbox'*, and the other *'bbox'* where object-level bounding box is used in both phases. We augmented the data using techniques like resizing, adding salt and pepper noise and blurring. Since our problem involved pose monitoring, we did not apply the most common augmentation strategies like flipping and rotation for pose experts as 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. We emphasize here that curation of the datasets into different poses, where an unambiguous and definite pose structure is present is not expensive as many pose aware datasets like the Pose Aware Person Dataset [22] are available. Modeling the pose is an important characteristic for reducing problem complexity and is much easier than the effort required in part based annotation which gives only a comparable accuracy.

	birds		cars		aircrafts	
	\bbox	bbox	\bbox	bbox	\bbox	bbox
Proposed	76.3	78.4	87.9	92.0	82.5	83.9
BCNN [15] \ft	80.1	81.3	83.9	-	78.4	-
BCNN [15] ft	84.1	85.1	91.3	-	84.1	-
BGL [23]	75.9	80.4	86.0	90.5	-	-
MixDCNN [14]	-	81.1	-	-	-	-
SCDA [24]	80.5	-	85.9	-	79.5	-

Table 3: Performance comparison with state-of-the-art on standard datasets. '\ft' denotes without finetuning.

4.1. Analysis and Characterization

Pose Experts v/s 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 images distinctly segregable based on their pose. A single network is supplied with all the pose-related data together during training while the pose experts specialize in specific poses. 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 for the training, 150 from each pose. The single net and the pose expert accuracies are reported in Table 1 and Table 2.

Shallow Pose Experts: Having highlighted the usefulness of an ensemble 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 an ensemble of shallow pose experts that can be trained efficiently even with extremely small datasets, and still perform at par or better than the deep network. 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 layers, this leads to a significant decrease in trainable parameters. Results are illustrated in Table 2.

4.2. Comparison with the State-of-the-Art

For comparisons in this section, we use Reduced VGG16 network for pose experts in the footwear dataset, and VGG16 network for the same in the benchmark datasets. Our architecture involves very few trainable parameters in comparison to the state-of-the-art such as BCNN [15] which has a high dimensional 512*512 bilinear vector, obtained after taking cross product. Gradient with respect to this layer is computationally expensive. Small parameter size is a critical requirement for a practical FGVC solution. Table 3 shows the comparison. For more details on experiments and the quality of

features learnt by our model, please refer to the supplementary material.

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 increases 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 model, we get a further increase of about $\sim 0.7\%$.

FGVC-Aircrafts Dataset The FGVC-Aircrafts dataset [25] consists of 10,000 images of aircrafts spanning 100 models. Images are divided into 2 poses: left facing and right facing. One can argue that complementary images from these 2 poses could have been generated by flipping the training data. However, our hypothesis 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. Experiments on this dataset showed the maximum improvement from the single network performance. This can be attributed to the minimal representation of aircrafts in ImageNet, which has been used for finetuning models in the state-of-the-art as well as ours. Our model outperforms the single network which gives an accuracy of 74.1% by nearly 8.5%. Bilinear CNN [15] gives an accuracy of 84.1% on the dataset. Our model is able to perform better than the SCDA approach [24]. When bounding boxes are used, the result from our model improves to 83.9% from 82.5%.

Stanford Cars Dataset Stanford Cars dataset [26] contains 16,185 images of 196 car categories. Images from this dataset were divided into 3 poses: front facing, side facing and back facing. Trends obtained for the cars dataset are similar to those obtained in the case of aircrafts. Our model again performs well on this dataset due to the pose structure in the data. We outdo the single VGG16 network (best performing single network with accuracy 79.8%) using our approach by nearly 8.1%. BGL [23] and SCDA [24] are both outperformed by a margin of about 2% each. Results with bounding box annotation are better by around 4% (at 92.0%), which we speculate are due to more background clutter here compared to any other dataset.

CUB200-2011 Dataset CUB200-2011 is a 200 bird species recognition dataset which contains 11,788 images. We segregate the dataset into 3 poses: front, left facing and right facing. On this dataset, we fall slightly short of the state-of-the-art accuracy as given by [15]. The primary reason for this seems to be the lack of rigidity or consistent poses in the birds dataset. The dataset contains images of birds with extreme variation in pose (e.g., flying birds, swimming birds), inconsistent pose (head and torso in opposite directions), and variation in 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 for the same reason. However, the proposed approach gives similar accuracy as the other state-of-the-art approaches: MixDCNN [14], BGL [23] and SCDA [24].

4.3. Application to Clothing Classification

We test our model on clothing classification using the DeepFashion Attribute Prediction dataset [27] with two protocols. DeepFashion has 50 category labels with bounding box annotations, which we use to crop out a particular class at a time. We segregate the cropped images into three poses: front, back and side automatically using image meta-data, and perform classification using our model as well as that of BCNN [15]. In the first protocol, we take the entire category set of 50 labels. BCNN performs at 53.4% whereas our ensemble gives 55.7% (Front: 61.8%, Back: 58.2%, Side: 50.5%). Our pose isolation model is able to give significant improvement in the side pose category classification which is a considerably difficult problem for the fashion domain. We also tested on a variant of the dataset where we grouped together visually similar clothing classes, to give a combined set of 19 classes. Details about the new class grouping can be found in the supplementary material. For this experiment, BCNN gave an accuracy of 74.5% while our pose ensemble performs at 79.6% (Front: 85.1%, Back: 81.1%, Side: 72.3%).

5. CONCLUSION

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, in a data scarce FGVC problem. The problem becomes even harder when the class has limited representation in large benchmark datasets like ImageNet, making it harder to pretrain for generic features.

We observe that the classification problem gets significantly simplified when viewing objects 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. In agreement with our hypothesis, we observe that the proposed approach improves the state-of-the-art by a greater margin as the category becomes more and more under-represented in ImageNet. The under-representation forces the models to learn new features from the relatively small number of fine-grained samples. The exploitation of structure in the data, such as pose, therefore becomes very important. We show that the proposed model excels in such cases.

6. REFERENCES

[1] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar, "Cats and dogs," in *CVPR*. IEEE, 2012, pp. 3498–3505.

[2] Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *CVPR*, 2014, pp. 2011–2018.

[3] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," Tech. Rep. CNS-TR-2011-001, Caltech, 2011.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. IEEE, 2009, pp. 248–255.

[5] Yuning Chai, Victor Lempitsky, and Andrew Zisserman, "Symbiotic segmentation and part localization for fine-grained categorization," in *ICCV*, 2013, pp. 321–328.

[6] Neeraj Kumar, Peter Belhumeur, Arijit Biswas, David Jacobs, WJWJ Kress, Ida Lopez, and João Soares, "Leafsnap: A computer vision system for automatic plant species identification," *ECCV*, pp. 502–516, 2012.

[7] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell, "Part-based R-CNNs for fine-grained category detection," in *ECCV*. Springer, 2014, pp. 834–849.

[8] Jia Deng, Jonathan Krause, and Li Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *CVPR*, 2013, pp. 580–587.

[9] Ryan Farrell, Om Oza, Ning Zhang, Vlad I Morariu, Trevor Darrell, and Larry S Davis, "Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance," in *ICCV*. IEEE, 2011, pp. 161–168.

[10] Ning Zhang, Ryan Farrell, and Trever Darrell, "Pose pooling kernels for sub-category recognition," in *CVPR*. IEEE, 2012, pp. 3665–3672.

[11] Thomas Berg and Peter N Belhumeur, "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *CVPR*, 2013, pp. 955–962.

[12] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.

[13] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik, "Deformable part models are convolutional neural networks," in *CVPR*, 2015, pp. 437–446.

[14] ZongYuan Ge, Alex Bewley, Christopher McCool, Peter Corke, Ben Upcroft, and Conrad Sanderson, "Fine-grained classification via mixture of deep convolutional neural networks," in *WACV*. IEEE, 2016, pp. 1–6.

[15] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji, "Bilinear CNN models for fine-grained visual recognition," in *ICCV*, 2015, pp. 1449–1457.

[16] Iacopo Masi, Stephen Rawls, Gérard Medioni, and Prem Natarajan, "Pose-aware face recognition in the wild," in *CVPR*, 2016, pp. 4838–4846.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[19] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[21] A. Yu and K. Grauman, "Fine-Grained Visual Comparisons with Local Learning," in *CVPR*, June 2014.

[22] Vijay Kumar, Anoop Namboodiri, Manohar Paluri, and CV Jawahar, "Pose-aware person recognition," *arXiv preprint arXiv:1705.10120*, 2017.

[23] Feng Zhou and Yuanqing Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *CVPR*, 2016, pp. 1124–1133.

[24] Xiu-Shen Wei, Jian-Hao Luo, Jianxin Wu, and Zhi-Hua Zhou, "Selective convolutional descriptor aggregation for fine-grained image retrieval," *IEEE Trans. on IP*, vol. 26, no. 6, pp. 2868–2881, 2017.

[25] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," Tech. Rep., 2013.

[26] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei, "3D object representations for fine-grained categorization," in *ICCV Workshop*, 2013, pp. 554–561.

[27] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations," in *CVPR*, 2016.

[28] Bolei Zhou *et al*., "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.

A. BENCHMARK DATASETS

Benchmark datasets used in the research were segregated pose-wise to suit the needs of the proposed model. CUB200-2011 is a bird species recognition dataset which was segregated into 3 poses: front facing, left facing and right facing. The FGVC-Aircrafts dataset has been divided into 2 poses: left facing and right facing, while the Stanford Cars dataset is divided into 3 poses: front facing, side facing and back facing. Representative images from these datasets are shown in Fig. 3 and Fig. 4.

Fig. 3: Representative images from FGVC-Aircrafts dataset [25] which was divided into two poses: left facing, right facing.

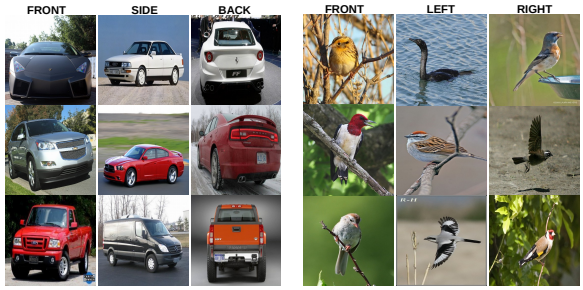

Fig. 4: (Left) Representative images from the Stanford Cars dataset [26] which was divided into three poses: front, side, back. (Right) Images from the CUB200-2011 birds dataset [3], divided into three poses: front, left facing, right facing.

B. FOOTWEAR DATASET

A contribution of this paper is the pose aware Footwear dataset of 1000 images, spanning 12 footwear classes. Representative image of our dataset is shown in Fig. 6.

C. COMPARISON WITH STATE-OF-THE-ART

For all the benchmark datasets, individual pose expert accuracies and pose detector performance, their ensemble comparison with the corresponding single network using VGG16, have been mentioned in Table 4. Fig. 8 shows the average precision-recall curves across the 4 datasets. Common mistakes made by our network are illustrated in Fig. 7, which is a visual comparison between top two pairs of most confused classes from each of the benchmark datasets used. Their respective confusion matrices are shown in Fig. 5.

	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. 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.

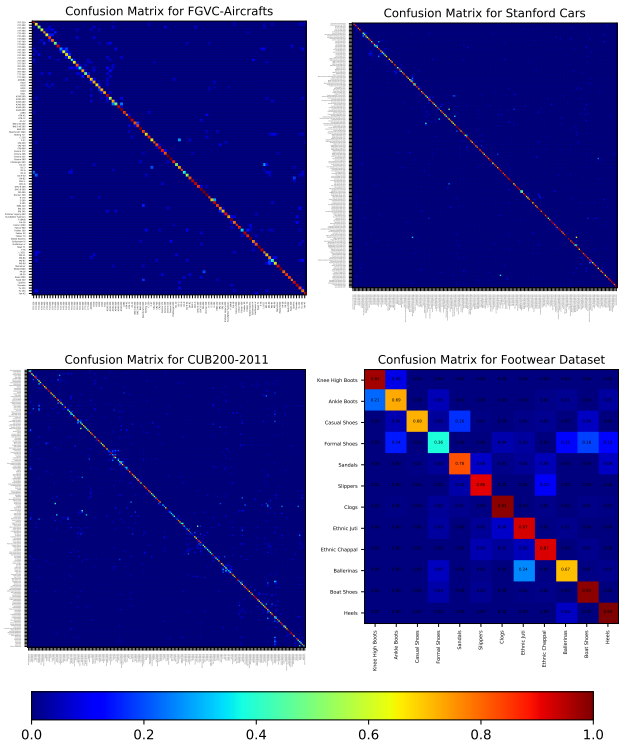

Fig. 5: Normalised confusion matrices for FGVC-Aircrafts, Stanford Cars, CUB200-2011 and our Footwear datasets. Please zoom in the pdf to read the fine text in the matrices.

Fig. 6: Representative images from the contributed footwear dataset. Please refer to the main paper for details of the dataset.

Fig. 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.

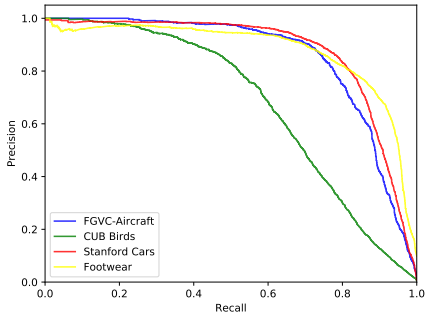

Fig. 8: Comparison of the precision-recall curves for the 4 datasets.

C.1. Application on Clothing Classification using DeepFashion

For a grouped category set of 19 classes from the DeepFashion Attribute Prediction dataset [27], category combinations are given in Table 5.

class	grouped categories
1	Parka, Anorak, Jacket, Bomber
2	Hoodie
3	Peacoat, Blazer, Coat
4	Sweater, Cardigan, Turtleneck
5	Button Down, Flannel
6	Henley, Tee, Top, Jersey, Blouse, Halter, Tank
7	Chinos
8	Culottes
9	Skirt
10	Cutoffs, Shorts, Sweatshorts, Trunks
11	Jeans, Jeggings, Capris
12	Joggers, Jodhpurs, Leggings
13	Sarongs
14	Gauchos
15	Caftan, Kaftan, Kimono, Cape, Coverup, Poncho
16	Jump-suit
17	Dress, Romper, Sundress, Shirtdress
18	Nightdress, Robe
19	Onesie

Table 5: Grouped category set for the experiment on 19 classes of the DeepFashion Attribute Prediction dataset.

C.2. Visualization

We use activation maps to highlight the quality of features being learnt by our networks, and help in visualizing how well the pose experts are able to localize the discriminative image regions which could vary in different poses of the same object. Fig. 9 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 the Stanford Cars dataset, the pose experts seem to focus on the front and hind wheels, backlight and roof in the first image

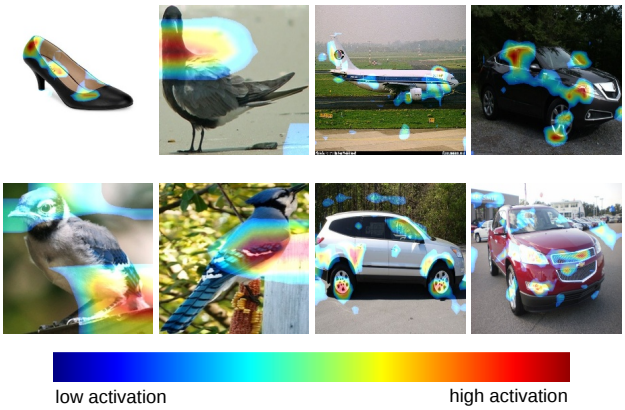

Fig. 9: Class Activation Maps. First row shows the activation maps from each of the 4 datasets. Second row shows 2 pairs of images, each pair belonging to a particular class, with different viewpoints and the variation in their discriminative regions. This discriminative information in different poses is directly used by our Pose Experts.

whereas the headlight and logo in the second image. These images have been generated using the technique suggested by Zhou *et al*. [28].