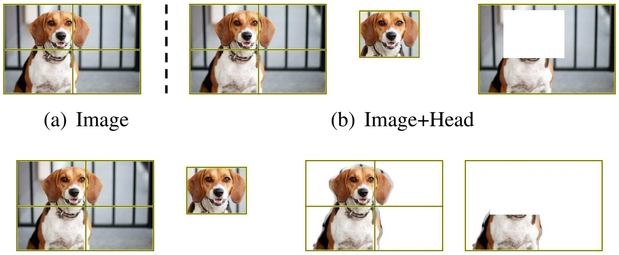

(a) Image

(b) Image+Head

(c) Image+Head+Body

Figure 4. **Spatial histogram layouts.** The three different spatial layouts used for computing the image descriptors. The image descriptor in each case is formed by concatenating the histograms computed on the individual spatial components of the layout. The spatial bins are denoted by yellow-black lines.

in correspondence of the pet head (as for the *image+head layout*) as well as other spatial bins computed on the foreground object region and its complement, as described next and in Fig. 4c. The foreground region is obtained either from the automatic segmentation of the pet body or from the ground-truth segmentation to obtain a best-case baseline. The foreground region is subdivided into five spatial bins, similar to the *image layout*. An additional bin obtained from the foreground region with the head region removed and no further spatial subdivisions is also used. Concatenating the histograms for all the spatial bins in this layout results in a 48,000 dimensional feature vector.

3.3. Automatic segmentation

The foreground (pet) and background regions needed for computing the appearance descriptors are obtained automatically using the grab-cut segmentation technique [36]. Initialization of grab-cut segmentations was done using cues from the over-segmentation of an image (*i.e*, superpixels) similar to the method of [15]. In this method, a SVM classifier is used to assign superpixels a confidence score. This confidence score is then used to assign superpixels to a foreground or background region to initialize the grab-cut iteration. We used Berkeley's ultrametric color map (UCM) [13] for obtaining the superpixels. Each superpixel was described by a feature vector comprising the color histogram and Sift-BoW histogram computed on it. Superpixels were assigned a score using a linear-SVM [21] which was trained on the features computed on the training data. After this initialization, grab-cut was used as in [34]. The improved initialization achieves segmentation accuracy of 65% this improving over our previous method [34] by 4% and is about 20% better than simply choosing all pixels as foreground (*i.e*, assuming the pet foreground entirely occupies the image). (Tab. 2). Example segmentations produced by our method on the Oxford-IIIT Pet data are shown in Fig. 5.

Method	Mean Segmentation Accuracy
All foreground	45%
Parkhi *et al.* [34]	61%
This paper	65%

Table 2. **Performance of segmentation schemes.** Segmentation accuracy computed as intersection over union of segmentation with ground truth.

Dataset	Mean Classification Accuracy
Oxford-IIIT Pet Dataset	38.45%
UCSD-Caltech Birds	6.91%
Oxford-Flowers102	53.71%

Table 3. **Fine grained classification baseline.** Mean classification accuracies obtained on three different datasets using the VLFeat-BoW classification code.

4. Experiments

The models are evaluated first on the task of discriminating the family of the pet (Sect. 4.1), then on the one of discriminating their breed given the family (Sect. 4.2), and finally discriminating both the family and the breed (Sect. 4.3). For the third task, both hierarchical classification (*i.e*, determining first the family and then the breed) and flat classification (*i.e*, determining the family and the breed simultaneously) are evaluated. Training uses the Oxford-IIIT Pet train and validation data and testing uses the Oxford-IIIT Pet test data. All these results are summarized in Tab. 4 and further results for pet family discrimination on the ASIRRA data are reported in Sect. 4.1. Failure cases are reported in Fig. 7.

Baseline. In order to compare the difficulty of the Oxford-IIIT Pet dataset to other Fine Grained Visual Categorization datasets, and also to provide a baseline for our breed classification task, we have run the publicly available VLFeat [40] BoW classification code over three datasets: Oxford Flowers 102 [33], UCSD-Caltech Birds [14], and Oxford-IIIT Pet dataset (note that this code is a faster successor to the VGG-MKL package [41] used on the UCSD-Caltech Birds dataset in [14]). The code employs a spatial pyramid [30], but does not use segmentation or salient parts. The results are given in Table 3.

4.1. Pet family discrimination

This section evaluates the different models on the task of discriminating the family of a pet (cat Vs dog classification).

Shape only. The maximum response of the cat face detector (Sect. 3.1) on an image is used as an image-level score