Appendix

A PASCAL Pre-Processing

Birds in PASCAL Parts are segmented into 14 parts : ‘beak’, ‘head’, ‘left eye’, ‘left foot’, ‘left leg’, ‘left wing’, ‘neck’, ‘right eye’, ‘right foot’, ‘right leg’, ‘right wing’, ‘tail’ and ‘torso’. We perform the following pre-processing steps to create a dataset for bird part segmentation:

- Crop individual objects from a full image based on annotated pixel-wise instance masks with a margin of 20 pixels on all sides
- Drop the cropped image if part of any other instance is present in the crop
- Check if the cropped instance is of bird class
- Filter out crops that have no part segmentation annotation
- Convert from 14 to 11 classes by dropping ’neck’ class and merging ’leg’/’feet’ classes due to inconsistencies in the annotated labels
- Filter out crops which are too small i.e. if either height or width ≤ 80 pixels
- Filter out crops which contains mostly the head of the bird by calculating the ratio of area of body to head. If ratio is ≤ 0.95, we drop the crop[2].

After these pre-processing steps we have 536 centered bird images. Using the standard split of PASCAL VOC, we separate out a training set of 271 images. If the crop originates from a train set image it goes in the training set of our dataset. We randomly split the rest of the images into validation and test sets containing 132 and 133 images respectively.

B Architectures

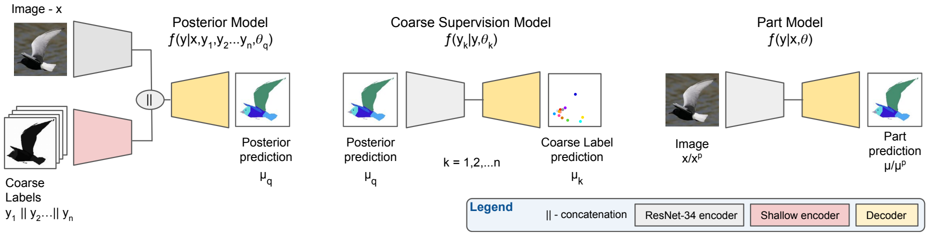

Fig. 5: Architectures of networks used in the EM method. This figure shows the input and output of each network. On the left is posterior model $f(y|x, y_{\text{mask}}, y_{\text{kp}})$, middle is $f(y_k|y)$ and right is $f(y|x)$. Notations are consistent with Alg. 1. Note that $f(y_{\text{mask}}|y)$ is deterministic, while $f(y_{\text{kp}}|y)$ is parameterized as a deep network and trained as part of the EM algorithm (see § 5.2).

[2] Note that we use this area ratio based method as the ’truncation’ and ’occlusion’ labels of PASCAL Part dataset would leave out many useful images

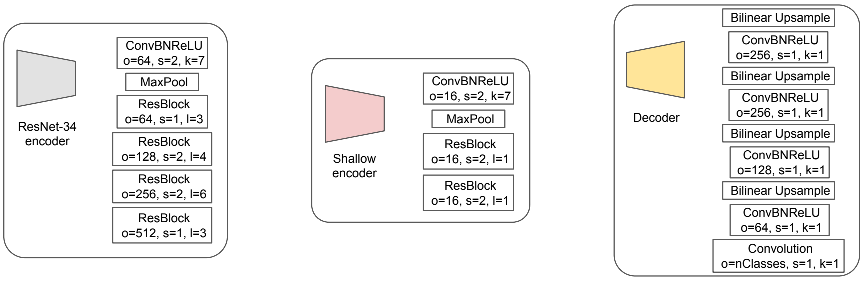

Fig. 6: Architectures in detail. Here we describe the model architectures used in the EM algorithm shown briefly in Fig. 5. ResBlock refers to the Basic Block used by ResNet architectures [11].

Note that the architectures used for the baselines are derivative of these architectures. The PointSup is made of the 'ResNet-34 encoder' and 'Decoder'. The MultiTask is made of one 'ResNet-34 encoder' and two 'Decoder's. The Finetuning and PsuedoSup baselines use 'ResNet-34 encoder' followed by 'Decoder'.

C Hyperparameters

Here we list the rest of the hyperparameters for EM training. The loss term for E Step is $\alpha\ell(\mu_q, \mu) + \lambda_1\ell_{\text{kp}}(y_{\text{kp}}, \mu_{\text{kp}}) + \lambda_2\ell_{\text{mask}}(y_{\text{mask}}, \mu_{\text{mask}}) + \gamma\ell_q(\mu_q)$. The loss for the part segmentation model is $\delta_1\ell(y^p, \mu^p) + \delta_2\ell(\mu_q, \mu)$. For ImageNet init, we set $\alpha = 0.05, \lambda_1 = 50, \lambda_2 = 1, \gamma = 0.01, \delta_1 = 0.05, \delta_2 = 50$. For keypoint init, $\alpha = 0.05, \lambda_1 = 100, \lambda_2 = 1, \gamma = 0.01, \delta_1 = 0.001, \delta_2 = 50$. For random init, $\alpha = 0.01, \lambda_1 = 50, \lambda_2 = 1, \gamma = 0.01, \delta_1 = 0.1, \delta_2 = 100$. For training on the Aircrafts dataset, we set $\alpha = 0.05, \lambda_2 = 5, \gamma = 0.01, \delta_1 = 0.01, \delta_2 = 50$. For tuning the hyperparameters the relative values of α and λ_1 (λ_2 for Aircrafts) is important. Similarly for δ_1 and δ_2. For α we sweep from [0.001,0.005,0.01,0.05,0.1,0.5], for λ_1: [10,50,100,500], for δ_1: [0.001,0.005,0.01,0.05,0.1,0.5], for δ_2: [10,50,100,500]. We keep γ as 0.01 for all cases.

D Results on PASCAL Val/Test Splits

In Table 5 we present the quantitative performance of models trained with ImageNet initialization on validation and testing sets of PASCAL Birds. A very similar trend as CUB dataset follows for PASCAL too.

Table 5: Results on PASCAL Val/Test.

Method	PASCAL Test	PASCAL Val
Finetuning	34.25	32.02
Multitask	30.34	29.50
PseudoSup	34.98	33.28
PointSup	35.72	35.78
Ours	36.31	37.52

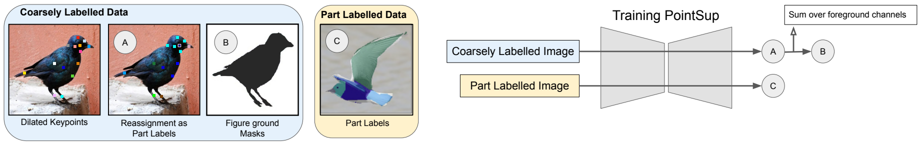

Fig. 7: Training PointSup using figure ground masks and handcrafted part labels for coarsely labelled data, and using part labels for part segmented data

E Qualitative Results

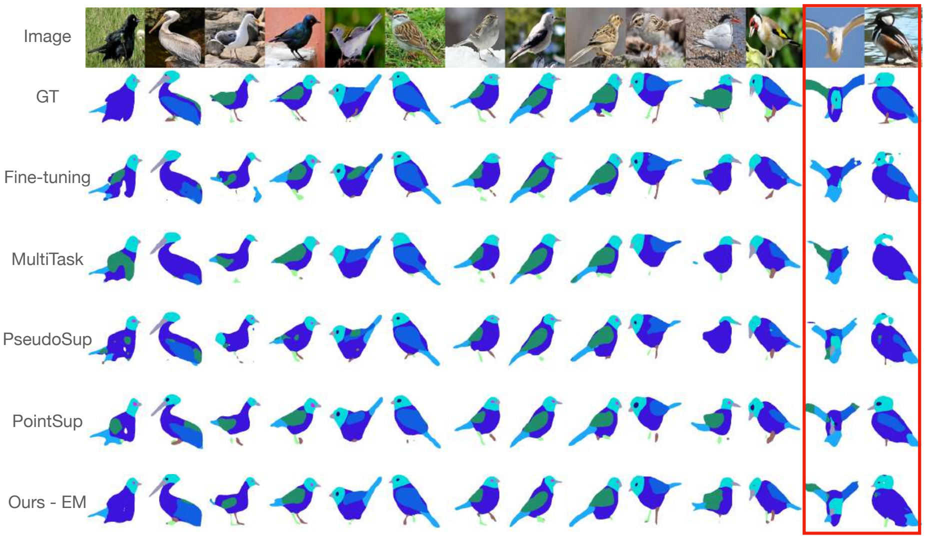

Fig. 8: Qualitative Results on PASCUB dataset. The examples bordered in red show two of the failure cases.

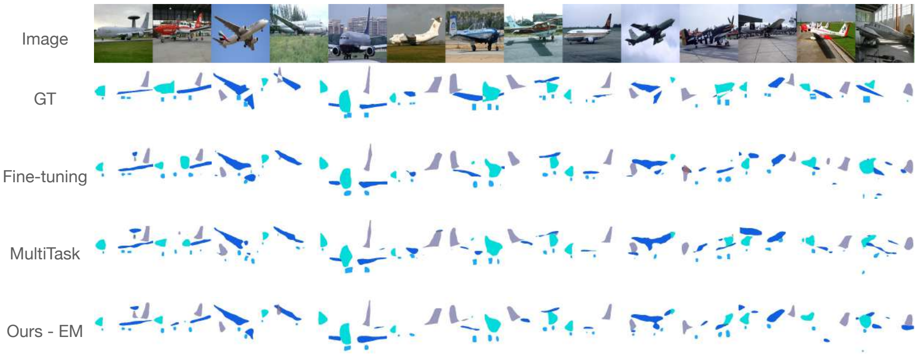

Fig. 9: Qualitative Results on OID Aircraft dataset.