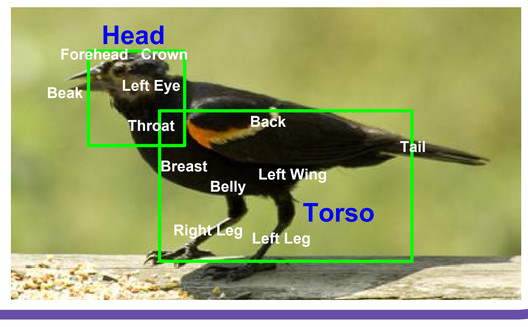
Part Localization using Multi-Proposal Consensus for Fine-Grained Categorization

Kevin J. Shih, Arun Mallya, Saurabh Singh, and Derek Hoiem
University of Illinois at Urbana-Champaign

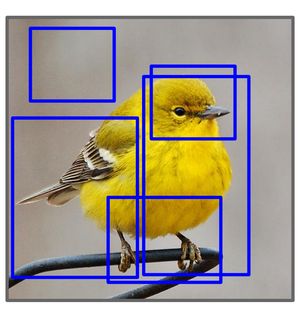
Task

Given an image of a bird, we wish to accurately determine not only where a keypoint is, but whether it is visible. With this information, we can use tight boxes around pre-defined sets of points to localize regions such as heads and torsos, which can in turn be used for part alignment for bird species classification.

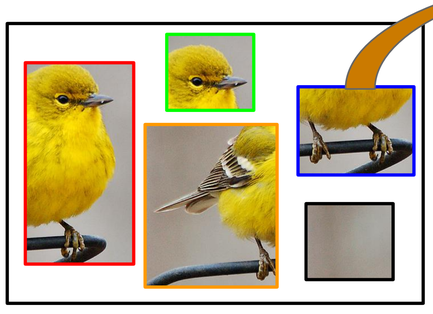


Training

We train a network to predict keypoint location and visibilities from image subcrops. We sample multiple EdgeBoxes and generate associated normalized keypoint locations and visibility masks. These are fed as training data into an AlexNet architecture with a multi-task loss. For improved performance, background-only crops are added as negatives.



Image



Sampled EdgeBoxes

Keypoint	$v, (l_x, l_y)$	$v, (l_x, l_y)$	$v, (l_x, l_y)$	
Beak	1,(0.5, 0.2)	1, (0.5, 0.6)	0,(DC,DC)	.
Breast	1,(0.5, 0.5)	0,(DC,DC)	0,(DC,DC)	.
Crown	1,(0.5, 0.1)	1, (0.3, 0.1)	0,(DC,DC)	.
Right Eye	1,(0.2, 0.2)	1, (0.3, 0.6)	0,(DC,DC)	.
Left Leg	1,(0.5, 0.8)	0,(DC,DC)	0,(DC,DC)	.
Right Leg	0,(DC,DC)	0,(DC,DC)	0,(DC,DC)	.
Tail	0,(DC,DC)	0,(DC,DC)	0,(DC,DC)	.

Label Information

AlexNet

Keypoint Visibilities v

$$\mathcal{L}_{vis} = ||v - \hat{v}||_2^2$$

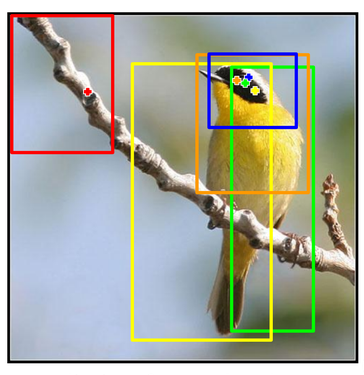
$$\mathcal{L}_{loc} = \sum_{i=1}^{N} v_i \cdot \left[(l_{x_i} - \hat{l}_{x_i})^2 + (l_{y_i} - \hat{l}_{y_i})^2\right]$$

$$\mathcal{L}_{net} = \mathcal{L}_{vis} + \mathcal{L}_{loc}$$

Keypoint Locations (l_x, l_y)

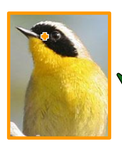
Inference

We sample EdgeBoxes and perform location and visibility predictions for each. Poor predictions are discarded with outlier removal.



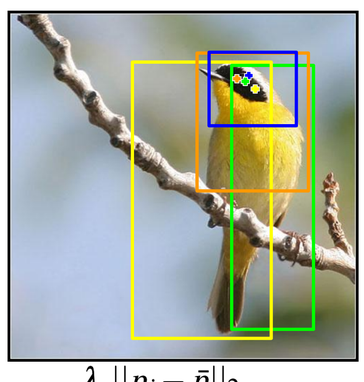
Prediction for each EdgeBox

0.1,(0.7,0.5)



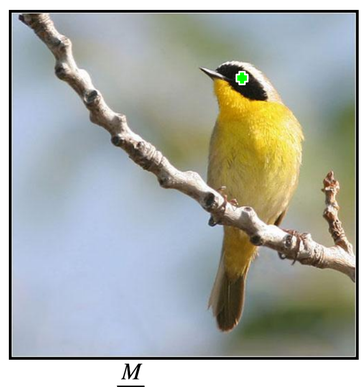
0.99,(0.3,0.2)

Drop predictions with low visibility confidences



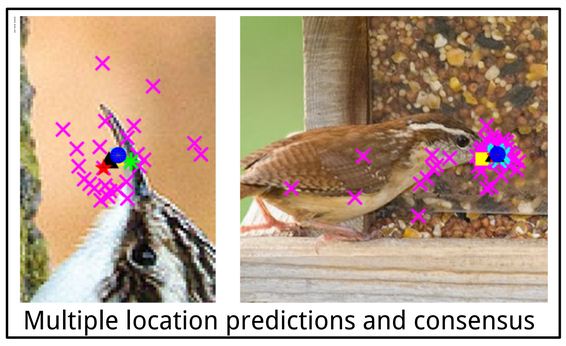
$$Z_i = \frac{\lambda\ ||p_i - \bar{p}||_2}{\text{median}\ (||p_i - \bar{p}||_2)},\ i \in \{1, M\}$$

Threshold by Z-score and return medoid

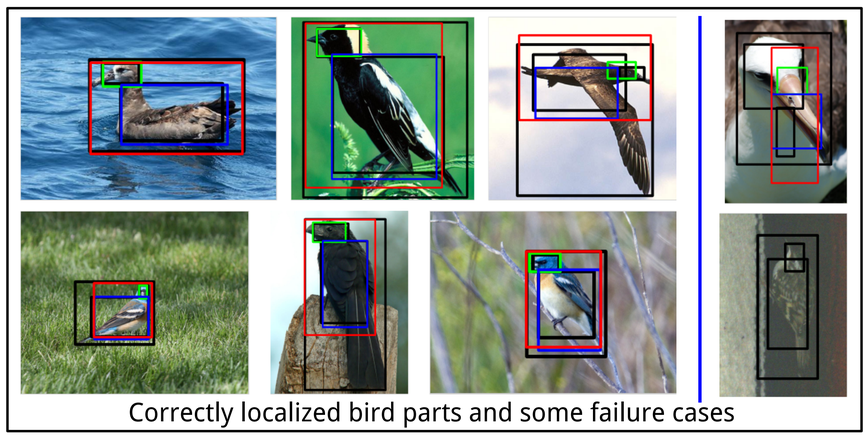


$$\bar{p} = \underset{p_j}{\text{argmin}} \sum_{i=1}^{M} ||p_j - p_i||_2,\ j \in \{1, M\}$$

Results and Analysis



Multiple location predictions and consensus



Correctly localized bird parts and some failure cases

Keypoint Localization & Visibility Accuracy

Method	PCP	AE	FVR	FIR
Poselets	24.5	2.89	47.9	17.2
Exemplar [1]	59.7	1.80	28.5	**4.5**
Ours	**69.1**	**1.4**	**17.1**	5.2

Part Localization Accuracy @ 50% IOU

	Method	Head	Torso	Body
GT Bbox	Part-Based RCNN [5]	68.2	79.8	N/A
	Deep LAC	74.0	**96.0**	N/A
	Ours (single GT box)	75.6	90.2	N/A
	Ours	**88.9**	94.3	N/A
No GT Bbox	Part-Based RCNN [5]	61.4	70.7	**88.3**
	Exemplar [1]	79.9	78.3	N/A
	Ours	**88.0**	**88.7**	84.6

Bird Classification Accuracy (200-way)

	Method	Accuracy
Oracle	Oracle Parts + SVM	81.5
GT Bbox	POOF	56.8
	Part-Based RCNN [5]	76.4
	Deep LAC	**80.3**
	Ours	**80.3**
No GT Bbox	Pose Norm	75.7
	Part-Based RCNN [5]	73.9
	Ours	**78.3**

We evaluate our model's localization and visibility accuracies with the standard metrics of PCP (Percent Correct Parts), AE (Average Error), FVR/FIR (False Visibility and Invisibility Rates) and demonstrate state-of-the-art results. The keypoint predictions are then used to localize head and torso parts, which we use for alignment in the species classification task. We thus demonstrate the good performance and usefulness of multiple predictions and consensus through a framework which is simpler than previous methods.

References

1. J. Liu and P. Belhumeur. Bird part localization using exemplar-based models with enforced pose and subcategory consistency. ICCV 2013
2. C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset.
3. C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. ECCV 2014
4. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. NIPS 2012
5. N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. ICCV 2014