Electrical Engineering & Computer Science

COLLEGE OF ENGINEERING

# Fine-Grained Recognition as HSnet Search for Informative Image Parts

Michael Lam, Behrooz Mahasseni and Sinisa Todorovic

Oregon State University

CVPR 2017

# Problem Statement: Fine-Grained Recognition

- **Given an image of an object, recognize its class**
- Categories are fine-grained and discriminated by subtle differences

Slaty Backed Gull

Western Gull

Slaty Backed Gull

Images from Caltech-UCSD Birds Dataset.

# Challenges of Fine-Grained Recognition

- **Different classes have similar appearance**
- Subtly differentiated by parts

Slaty Backed Gull

Western Gull

Slaty Backed Gull

Images from Caltech-UCSD Birds Dataset.

# Challenges of Fine-Grained Recognition

- **Same classes have different appearance**
- Variations in gender, season, location

Slaty Backed Gull

Western Gull

Slaty Backed Gull

Images from Caltech-UCSD Birds Dataset.

# Challenges of Fine-Grained Recognition

- **Variations in pose, viewpoint, background, lighting**


Slaty Backed Gull

Images from Caltech-UCSD Birds Dataset.

# Challenges of Fine-Grained Recognition

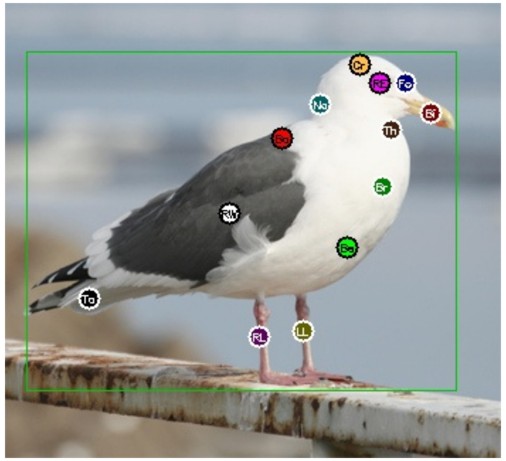
- **Background clutter: remaining image context outside of informative image parts may hurt recognition**

Ovenbird

Images from Caltech-UCSD Birds Dataset.

# Challenges of Fine-Grained Recognition

- **Small datasets, difficult if not impossible to obtain more data**
- E.g. biological datasets, military datasets



Slaty Backed Gull

Slaty Backed Gull

Western Gull

Images from Caltech-UCSD Birds Dataset.

# Prior Work: Fine-Grained Recognition

## Part-Based Models

- Localize parts and compare corresponding locations
- Factor out variations due to pose, viewpoint and location

- Farrell et al. 2011
- Zhang et al. 2014
- Branson et al. 2014
- …

- **Advantages**: High accuracy, factors out variations
- **Challenges**: Slow, part annotations required


1. Farrell et al. Birdlets: Subordinate Categorization using Volumetric Primitives and Pose-normalized Appearance. ICCV, 2011.
2. Zhang et al. Part-based R-CNNs for Fine-grained Category Detection. ECCV, 2014.
3. Branson et al. Bird Species Categorization Using Pose Normalized Deep Convolutional Nets. BMVC, 2014.

# Prior Work: Fine-Grained Recognition

## General Image Classification

- Just classify, no part annotations needed
- Modern approaches use CNN

- Jaderberg et al. 2015
- Lin et al. 2015
- …

- **Advantages**: Fast, does not require part annotations
- **Challenges**: Lower accuracy without parts information

1. Jaderberg et al. Spatial Transformer Networks. NIPS 2015.
2. Lin et al. Bilinear CNN Models for Fine-grained Visual Recognition. ICCV 2015.

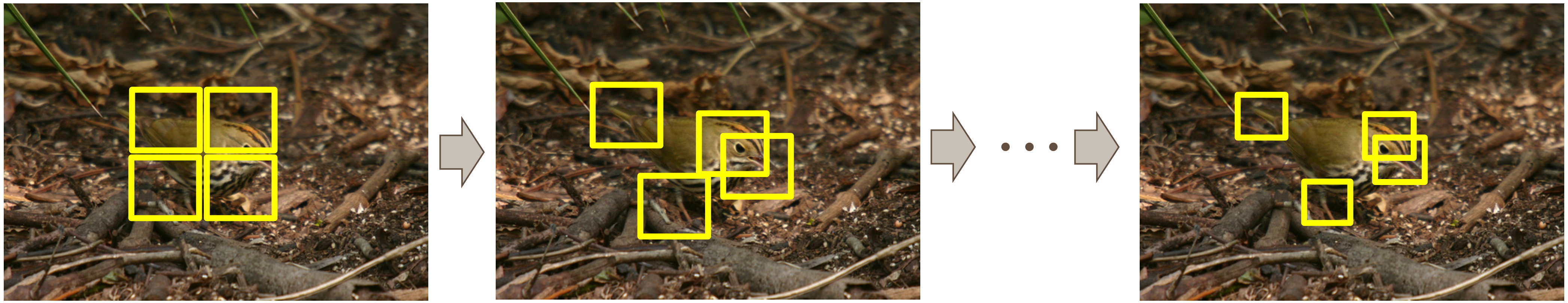
# Our Key Ideas

- **Part-based**: unlike object recognition, fine-grained recognition can benefit from removing background context and focusing on parts
- **Iterative**: instead of one shot reasoning, iteratively search for discriminative parts as bounding boxes in the image
- **Supervised and weakly supervised**: search for parts even without part annotations

# Our Approach

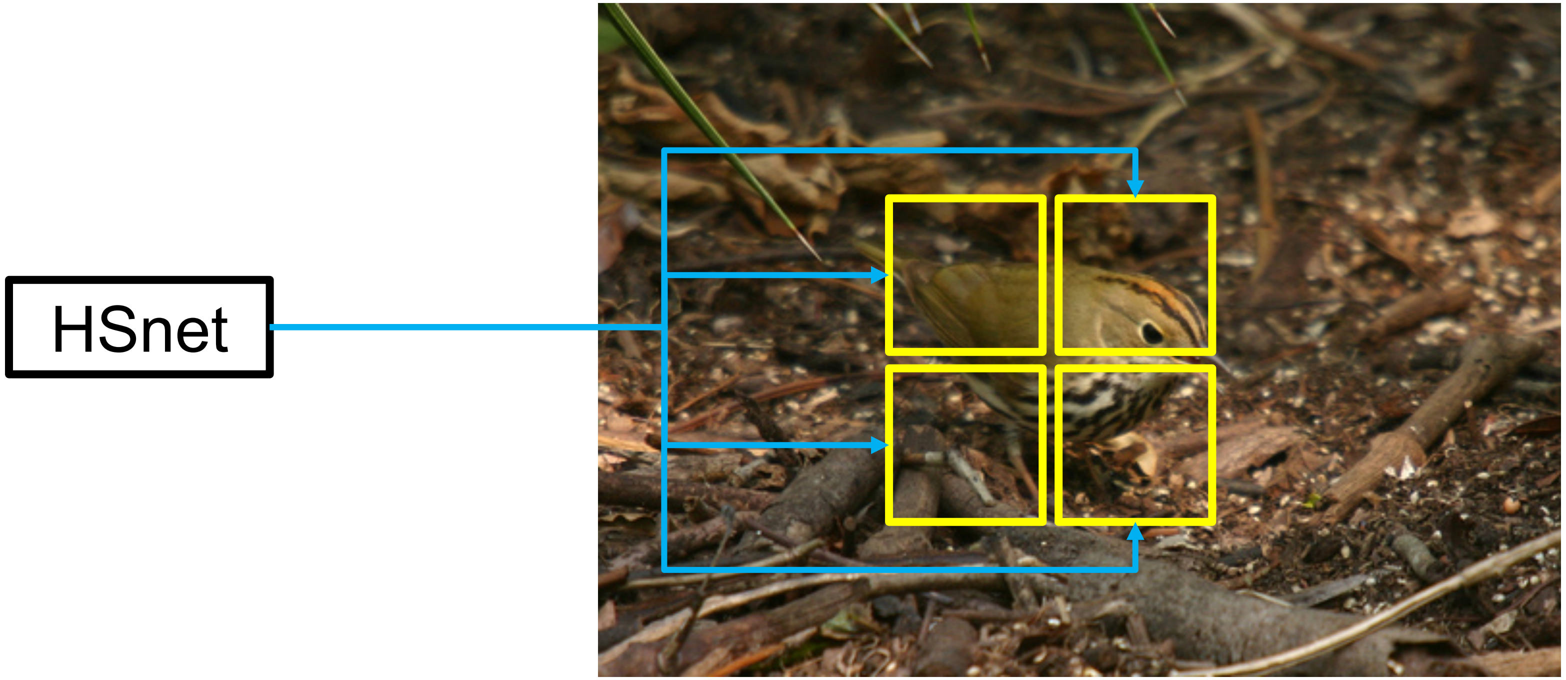
- Iterative approach for parts localization and class prediction

- In each iteration, improve localization and predict class
  - Localization and classification is guided by HSnet
  - Number of parts is fixed
  - Final iteration yields best localization and class prediction



Iteration 1 Iteration 2 Iteration $\tau$

# Our Approach

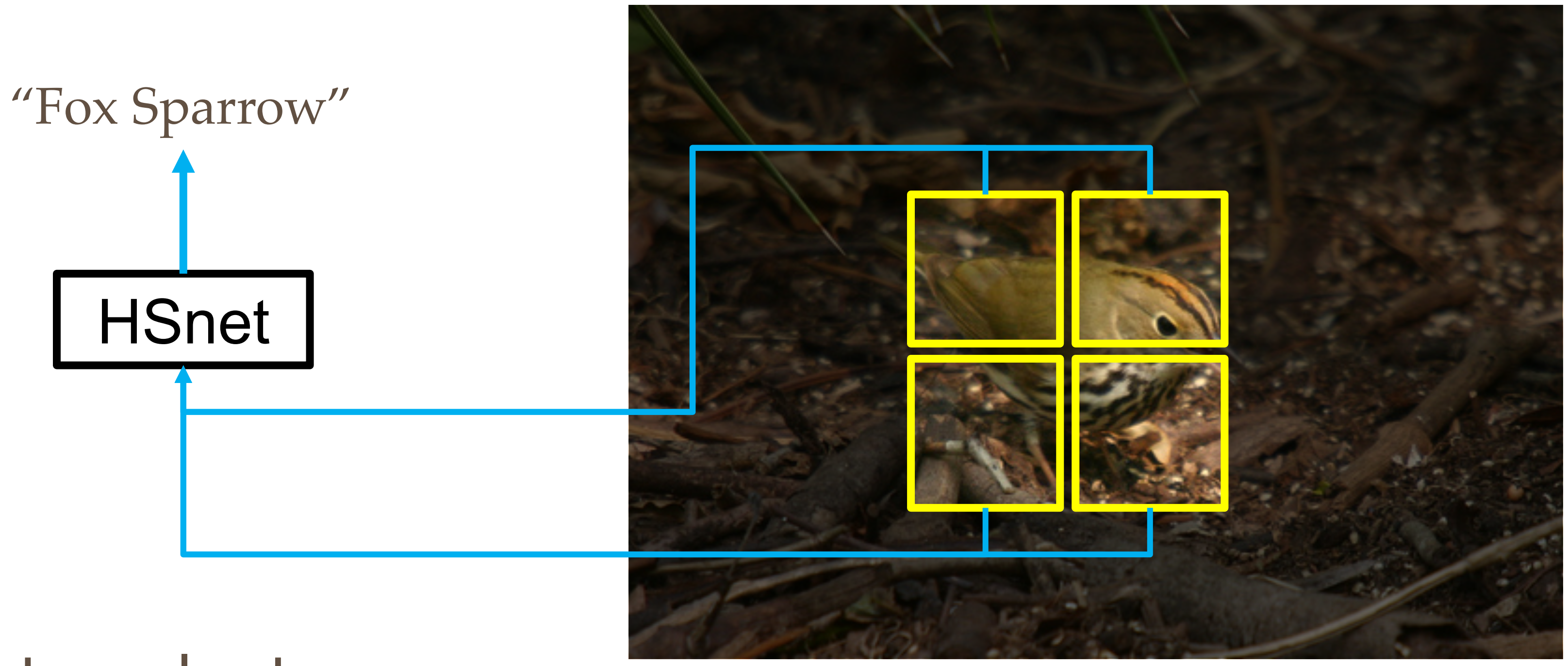



HSnet proposes initial bounding boxes

Iteration 1

(4 parts here for illustration purposes)

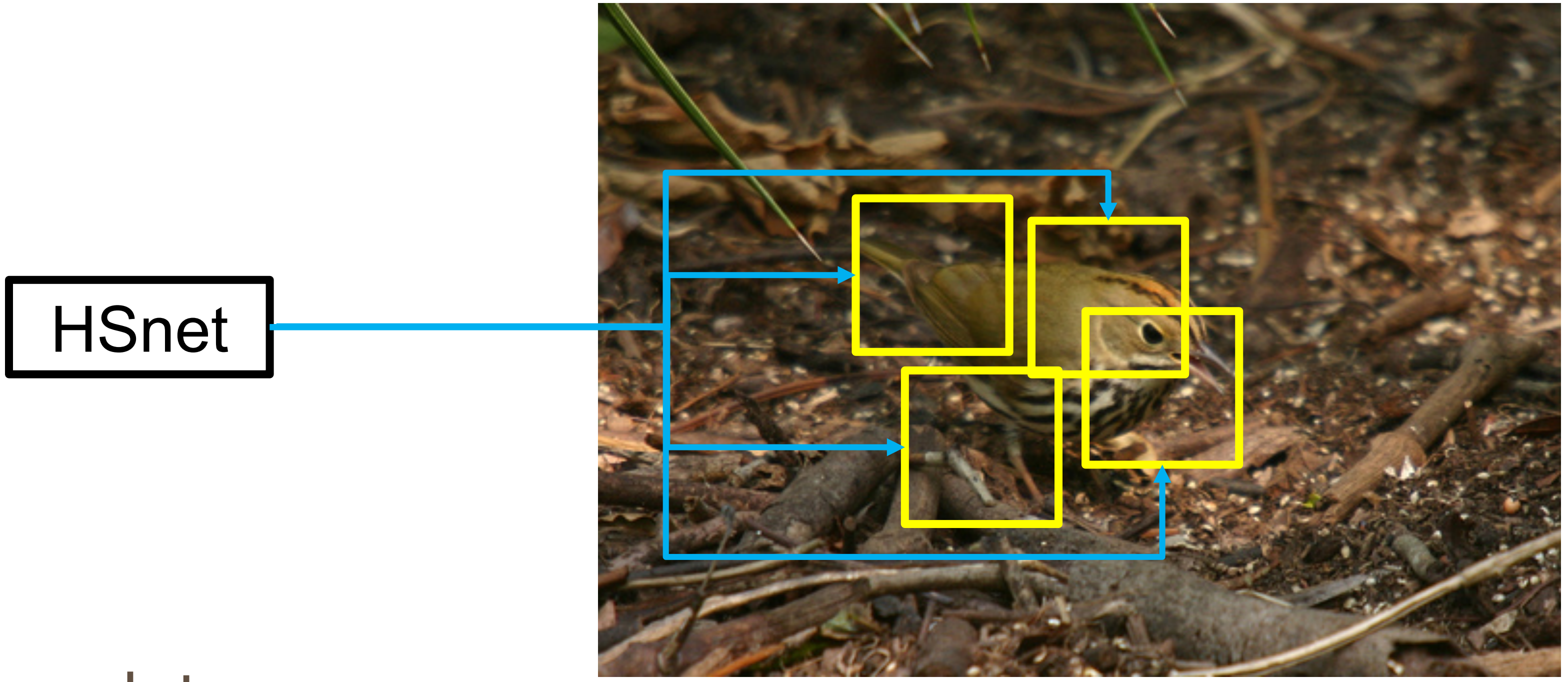
# Our Approach




Iteration 1

HSnet evaluates proposals for classification

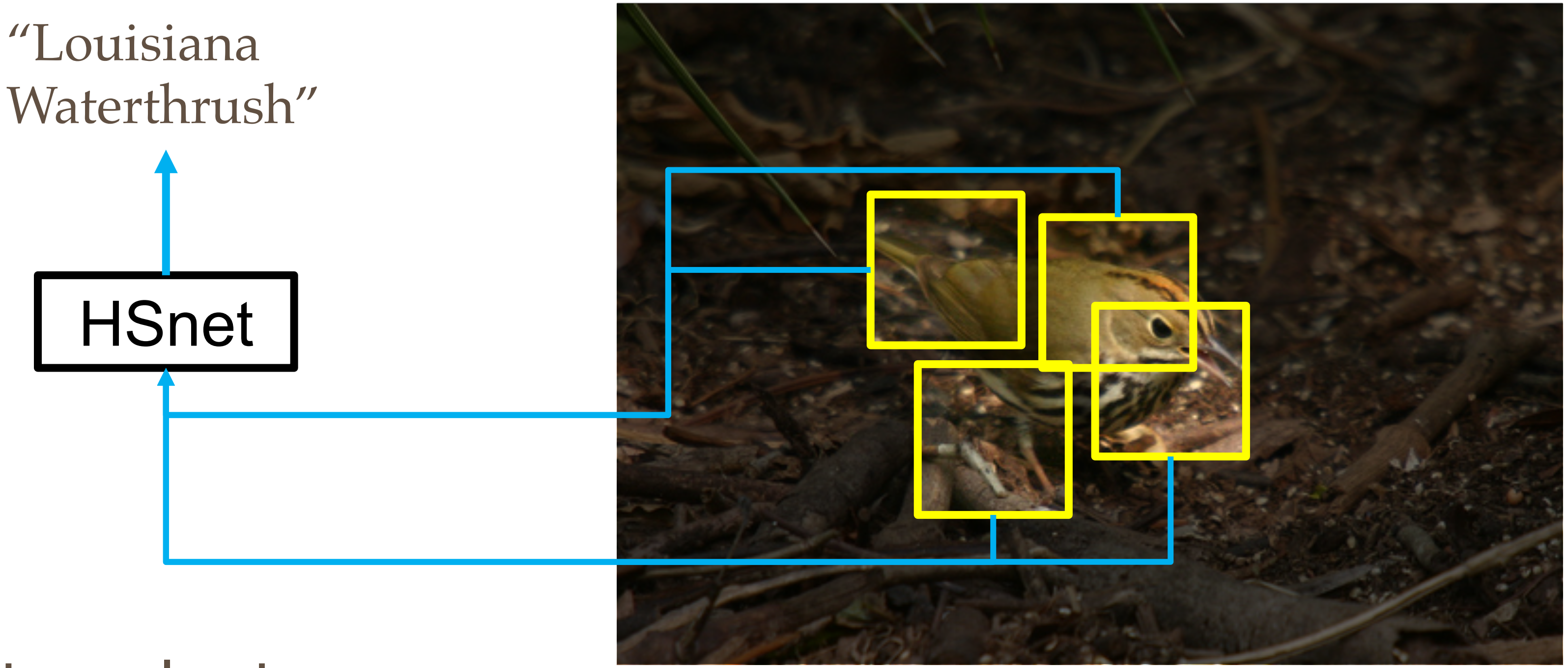
# Our Approach




Iteration 2

HSnet updates proposals

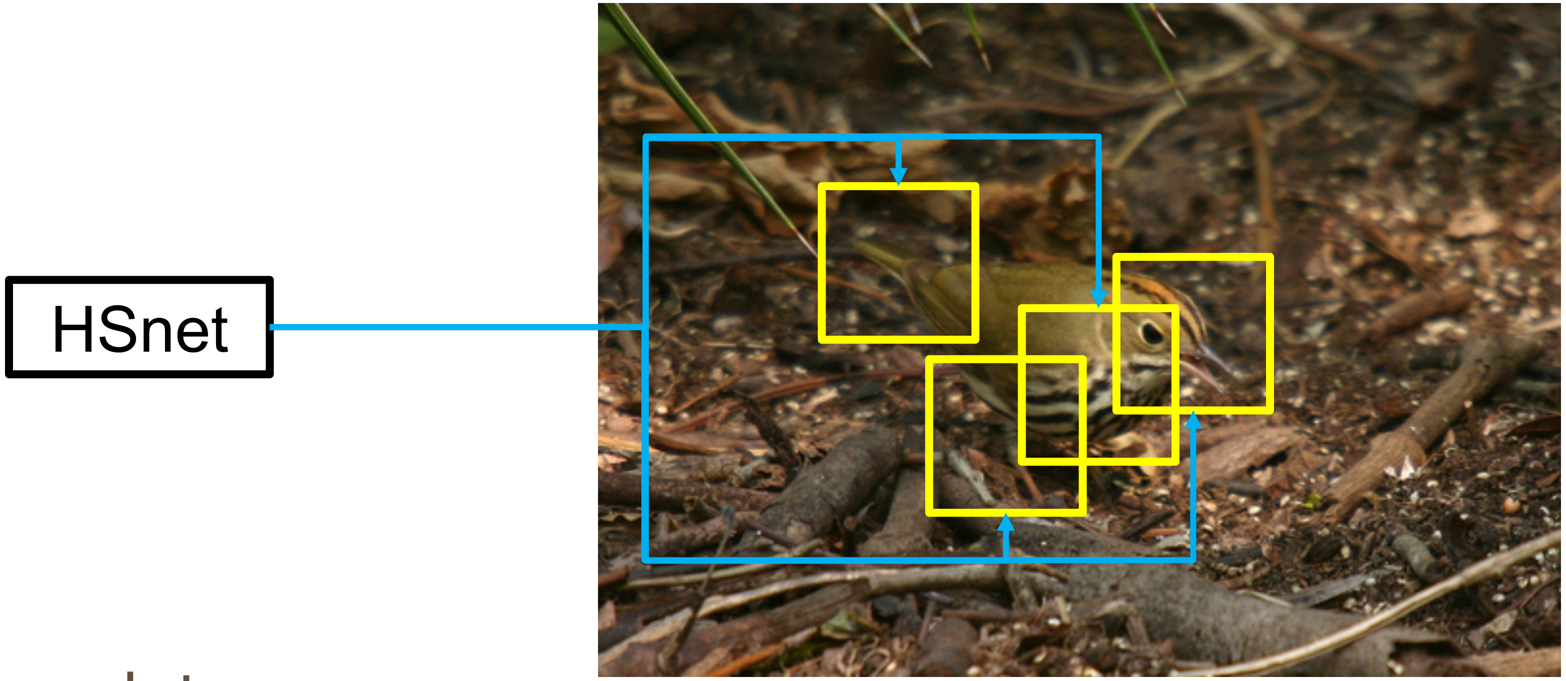
# Our Approach




HSnet evaluates proposals for classification

Iteration 2

# Our Approach




HSnet updates proposals

Iteration $\tau$

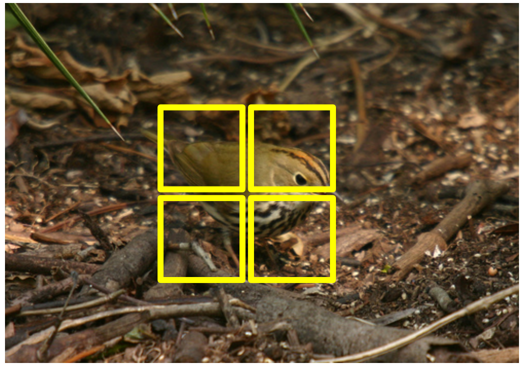
# Our Approach

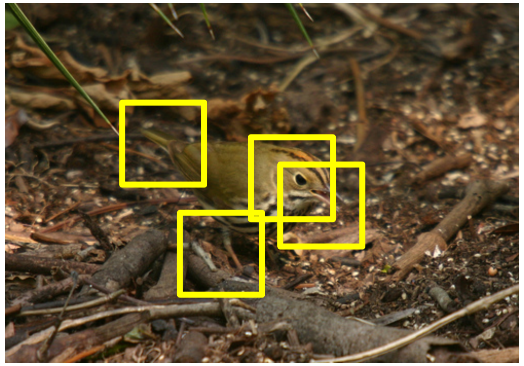

Iteration $\tau$

HSnet evaluates proposals for classification

# Search Formulation

- **State**: history of location and sizes of bounding box proposals





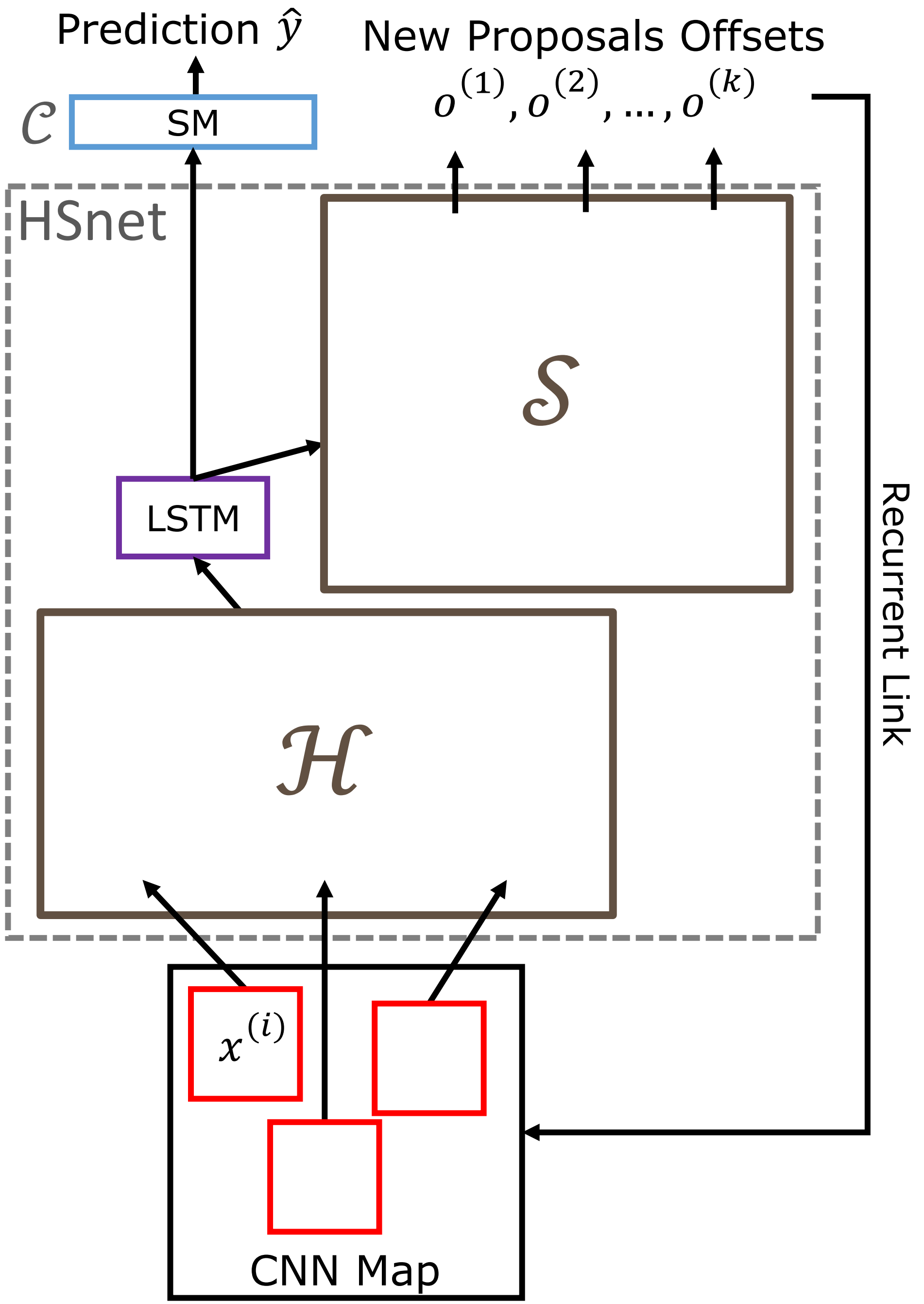
- **Heuristic function**: evaluates bounding box proposals
- **Successor function**: generates bounding box proposals
- **H**euristic and **S**uccessor functions are formulated as **HS**net

# HSnet Architecture

HSnet

# HSnet Architecture

- Heuristic $\mathcal{H}$ evaluates current state
- LSTM updates search history
- Classifier $\mathcal{C}$ makes prediction $\hat{y}$
- Successor $\mathcal{S}$ proposes candidate bounding boxes based on history

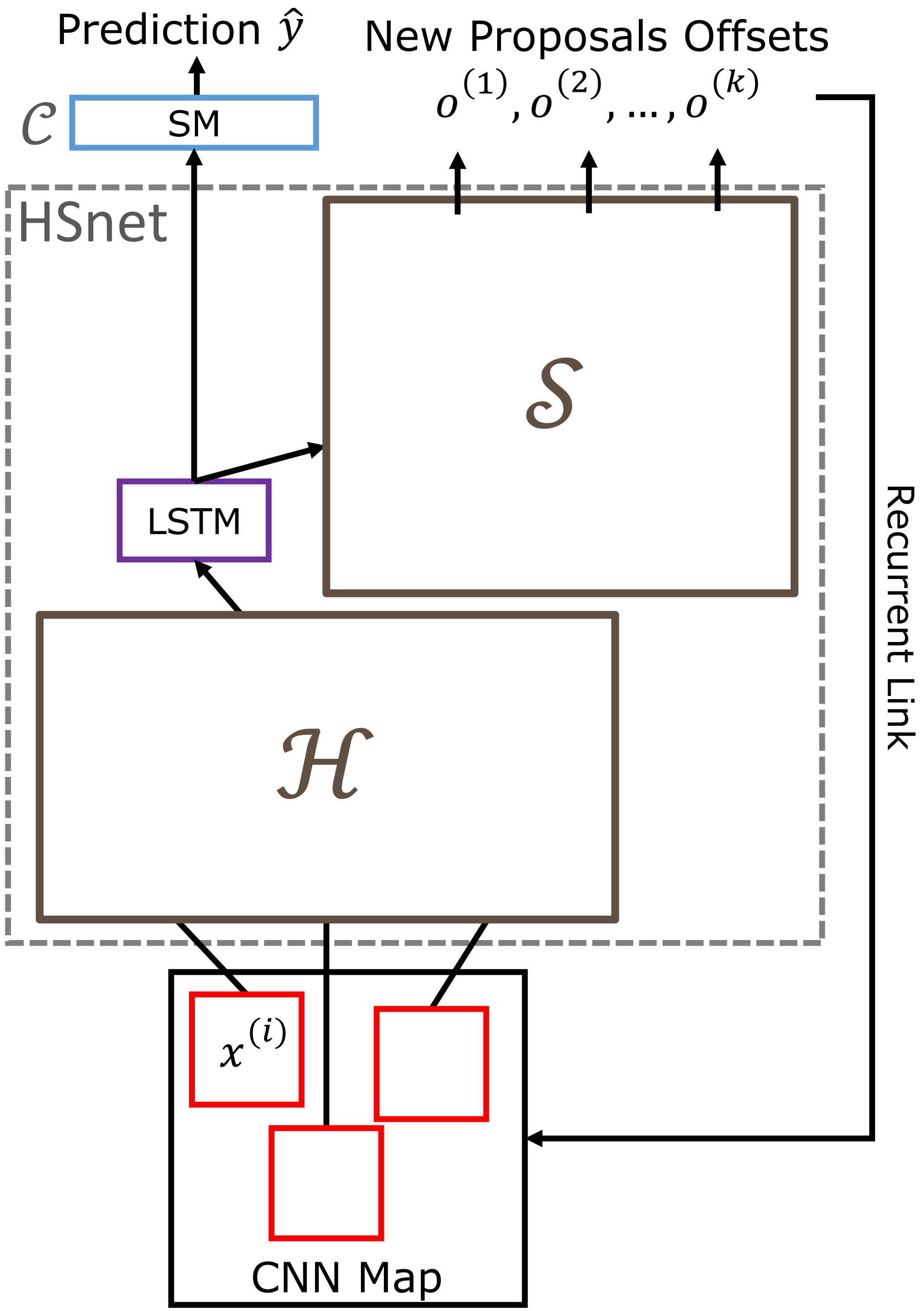



SM: Softmax

$x^{(i)}$: bounding box $i$ features
$o^{(i)}$: bounding box $i$ offset

# HSnet Architecture

- Heuristic $\mathcal{H}$ evaluates current state
- LSTM updates search history
- Classifier $\mathcal{C}$ makes prediction $\hat{y}$
- Successor $\mathcal{S}$ proposes candidate bounding boxes based on history

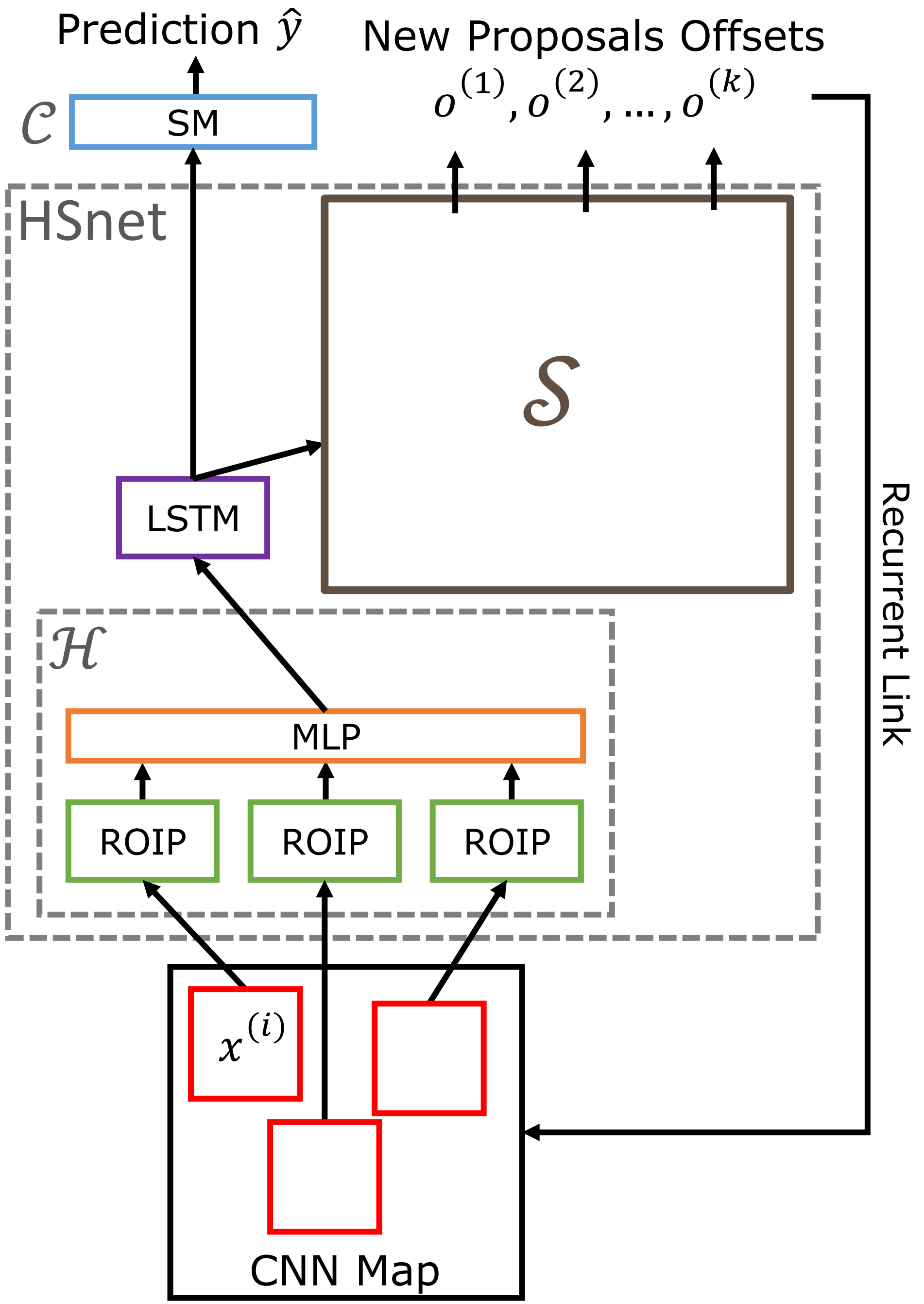



SM: Softmax
ROIP: Region of Interest Pooling
MLP: Multilayer Perceptron

$x^{(i)}$: bounding box $i$ features
$o^{(i)}$: bounding box $i$ offset

# HSnet Architecture

- Heuristic $\mathcal{H}$ evaluates current state
- LSTM updates search history
- Classifier $\mathcal{C}$ makes prediction $\hat{y}$
- Successor $\mathcal{S}$ proposes candidate bounding boxes based on history

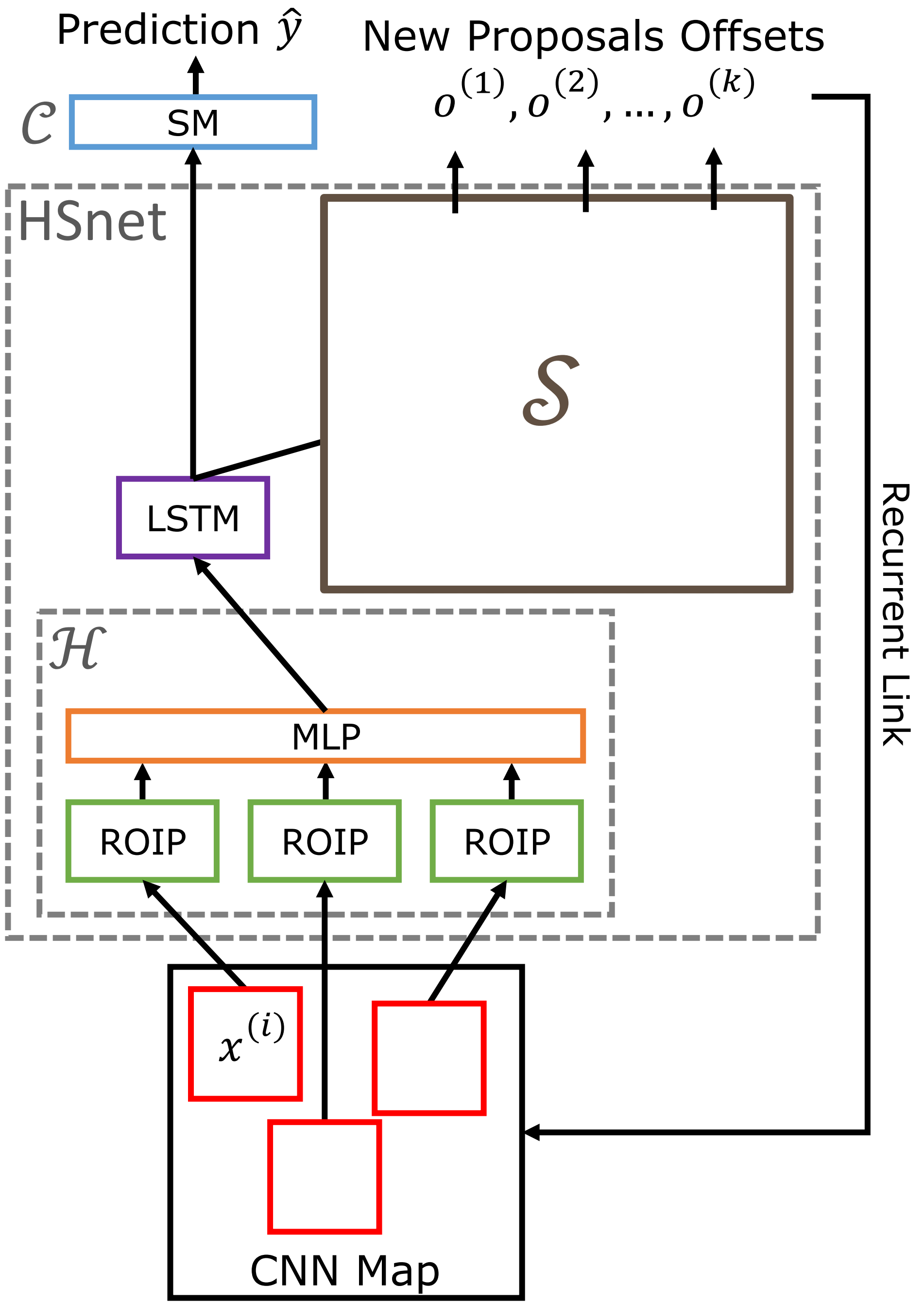



SM: Softmax
ROIP: Region of Interest Pooling
MLP: Multilayer Perceptron

$x^{(i)}$: bounding box $i$ features
$o^{(i)}$: bounding box $i$ offset

# HSnet Architecture

- Heuristic $\mathcal{H}$ evaluates current state
- LSTM updates search history
- Classifier $\mathcal{C}$ makes prediction $\hat{y}$
- Successor $\mathcal{S}$ proposes candidate bounding boxes based on history

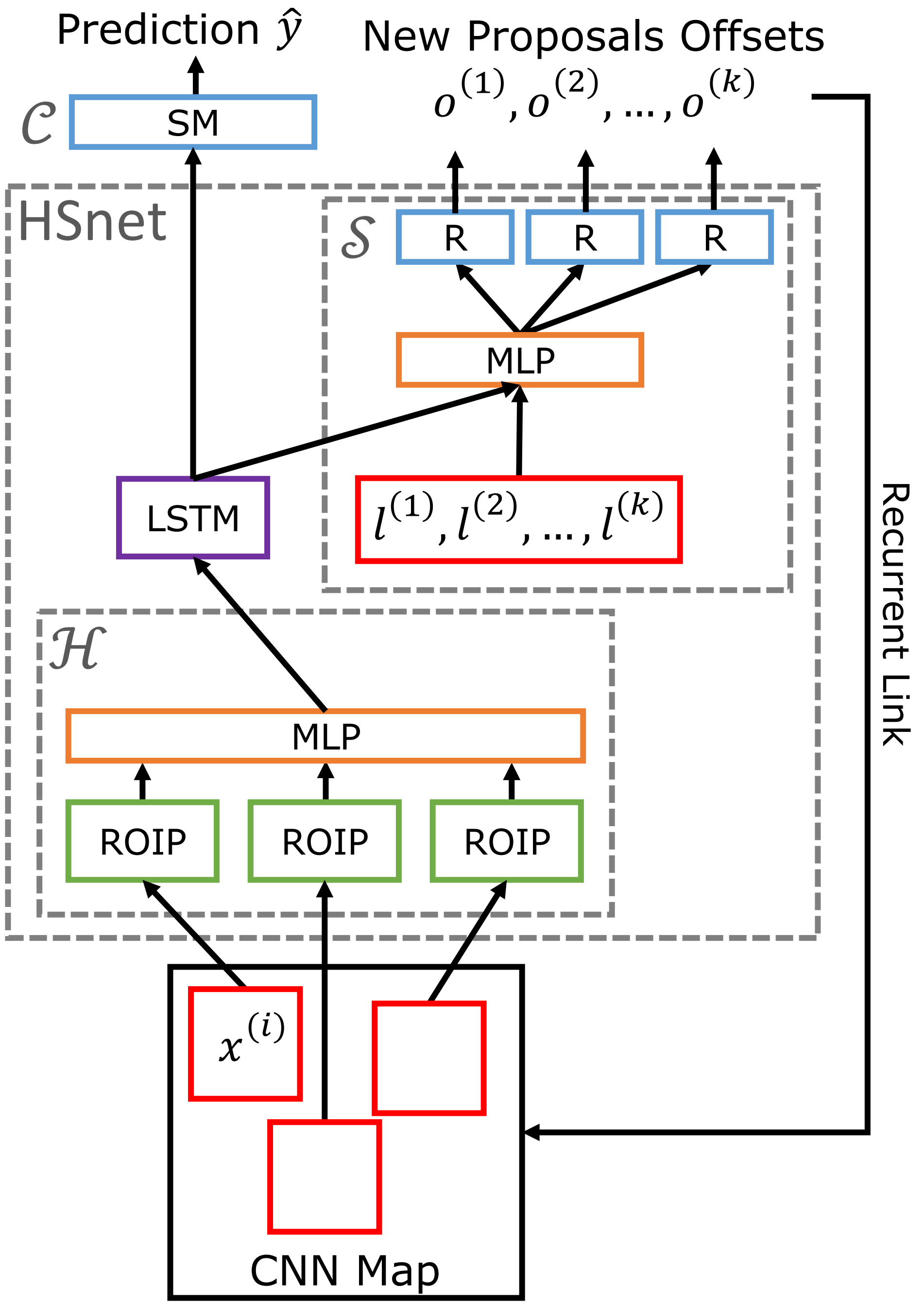



SM: Softmax
ROIP: Region of Interest Pooling
MLP: Multilayer Perceptron
R: Regression

$x^{(i)}$: bounding box $i$ features
$o^{(i)}$: bounding box $i$ offset
$l^{(i)}$: bounding box $i$ location

# HSnet Architecture

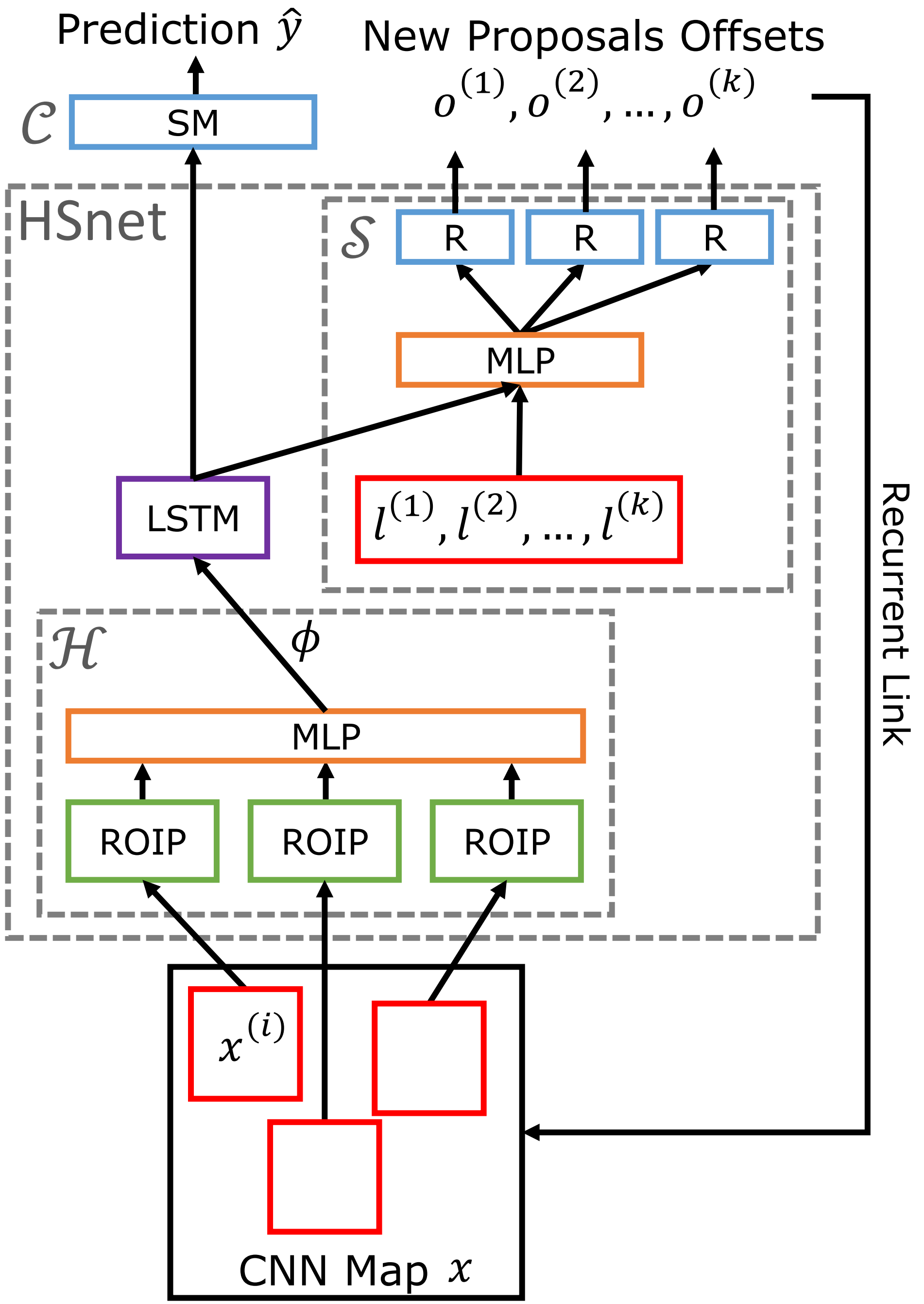
- Heuristic $\mathcal{H}$ evaluates current state
- LSTM updates search history
- Classifier $\mathcal{C}$ makes prediction $\hat{y}$
- Successor $\mathcal{S}$ proposes candidate bounding boxes based on history




SM: Softmax
ROIP: Region of Interest Pooling
MLP: Multilayer Perceptron
R: Regression

$x^{(i)}$: bounding box $i$ features
$o^{(i)}$: bounding box $i$ offset
$l^{(i)}$: bounding box $i$ location

# HSnet Architecture



# **Supervised** vs. Weakly Supervised

- When part annotations are available:

$$L = \underbrace{-\log p(y)}_{\text{Classification Loss}} + \underbrace{\sum_{t=1}^{\tau} \lambda_t \sum_{i=1}^{k} \left\| l^{(i)} - \hat{l}_t^{(l)} \right\|^2}_{\text{Parts Location Loss}}$$

$-\log p(y)$: cross entropy loss
$l^{(i)}$: groundtruth bounding box $i$ location
$\hat{l}_t^{(i)}$: predicted bounding box $i$ location at time $t$
$\lambda_t$: regularization parameter
$\tau$: time bound parameter
$k$: number of parts

# Supervised vs. **Weakly Supervised**

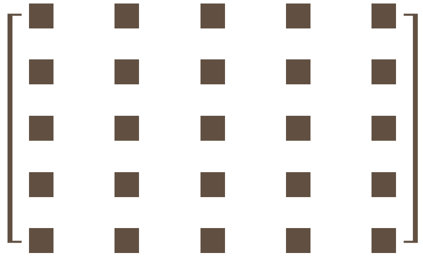
- When part annotations are not available:

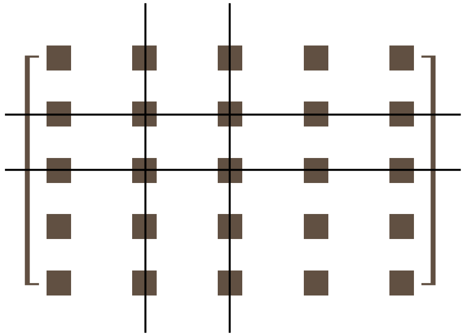
$$L = \underbrace{-\log p(y)}_{\text{Classification Loss}} - \underbrace{\sum_{t=1}^{\tau} \lambda_t \log P_t}_{\substack{\text{Diversity Regularization} \\ \text{(Determinantal Point Process)}}}$$

$$P_t = \frac{\det |\Omega_k|}{\det |\Omega + I|}$$

Probability of having diverse bounding box candidates at time $t$

$-\log p(y)$: cross entropy loss
$\lambda_t$: regularization parameter
$\tau$: time bound parameter
$k$: number of parts
$\Omega$: matrix of affinities between all possible bounding boxes
$\Omega_k$: restriction of $\Omega$ to $k$ selected bounding boxes

$\Omega =$ 

$\Omega_\mathrm{k} =$ 

# Datasets

**Caltech-UCSD Birds 200-2011**

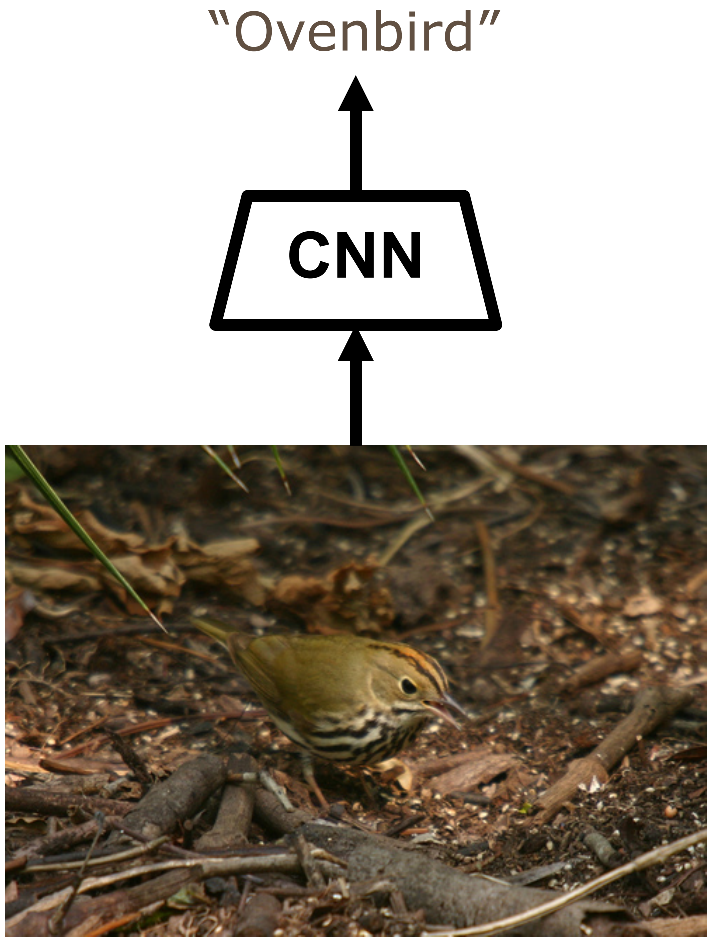
**Stanford Cars 196**

# Annotations

- Caltech UCSD Birds
  - Part locations provided, but no bounding box for each part
  - 15 parts: back, belly, bill, breast, crown, left eye, right eye, forehead, left leg, right leg, nape, tail, throat, left wing, right wing
- Stanford Cars
  - No parts annotation

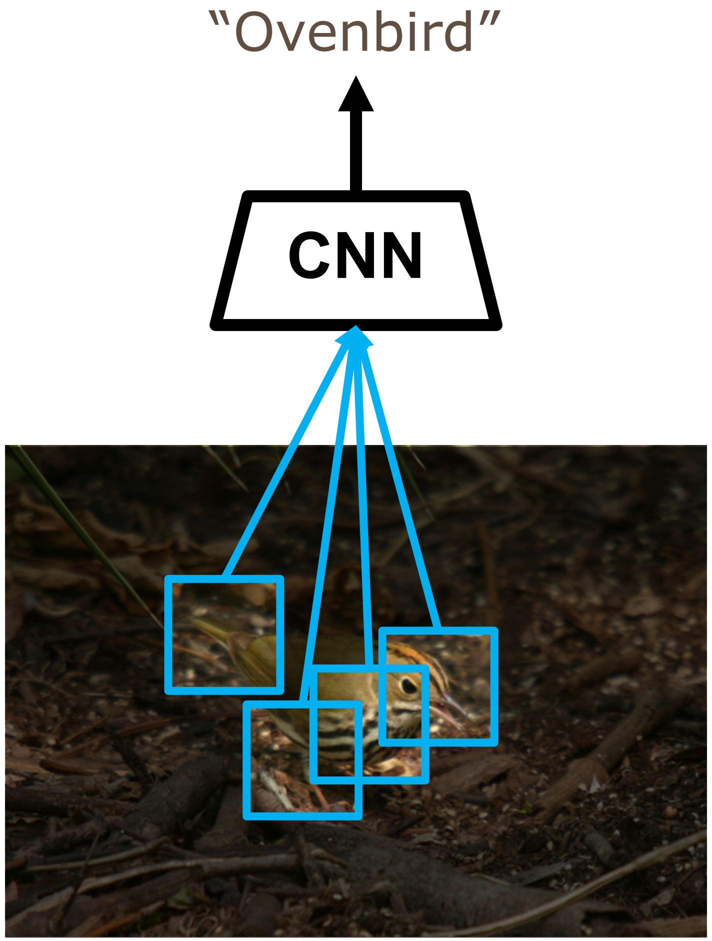
# Baselines

- **B1: CNN (fine-tuned)**
- B2: CNN with ground truth bounding boxes
- B3: HSnet with one ground truth bounding box
- B4: HSnet with one bounding box



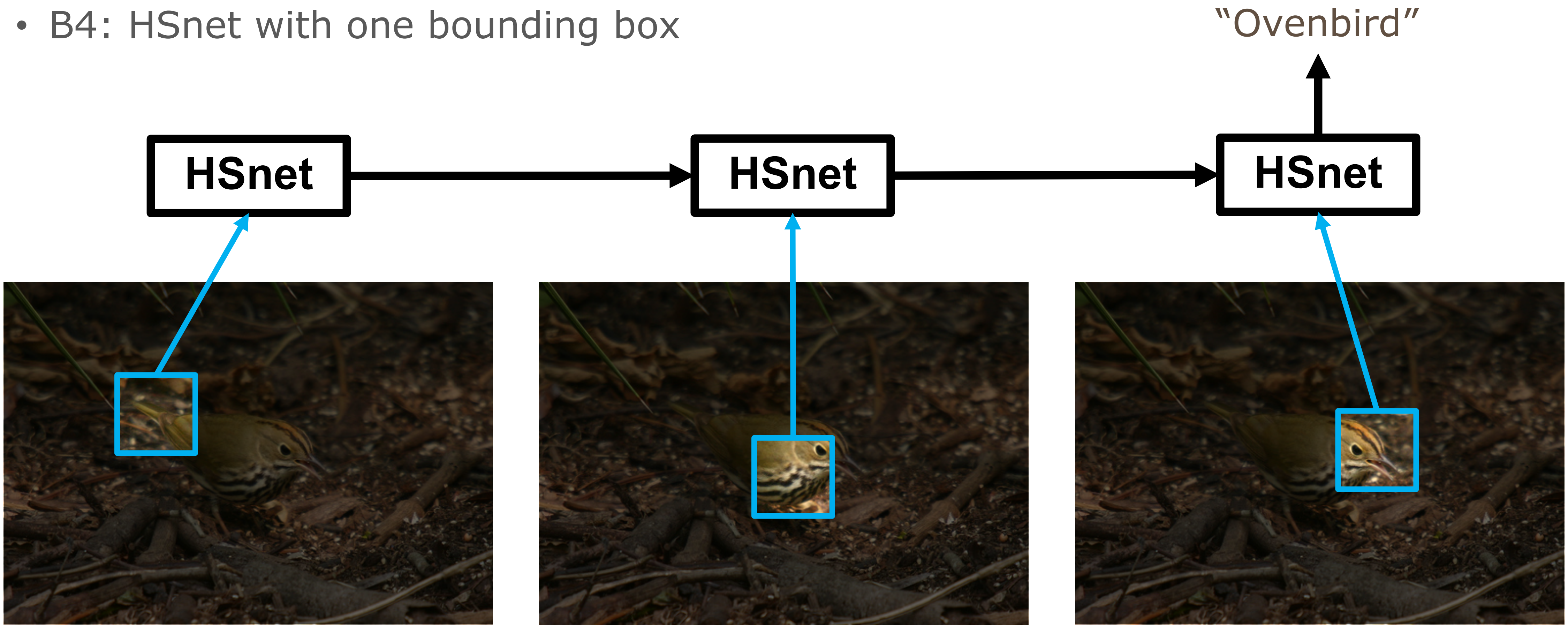
# Baselines

- B1: CNN (fine-tuned)
- **B2: CNN with ground truth bounding boxes**
- B3: HSnet with one ground truth bounding box
- B4: HSnet with one bounding box



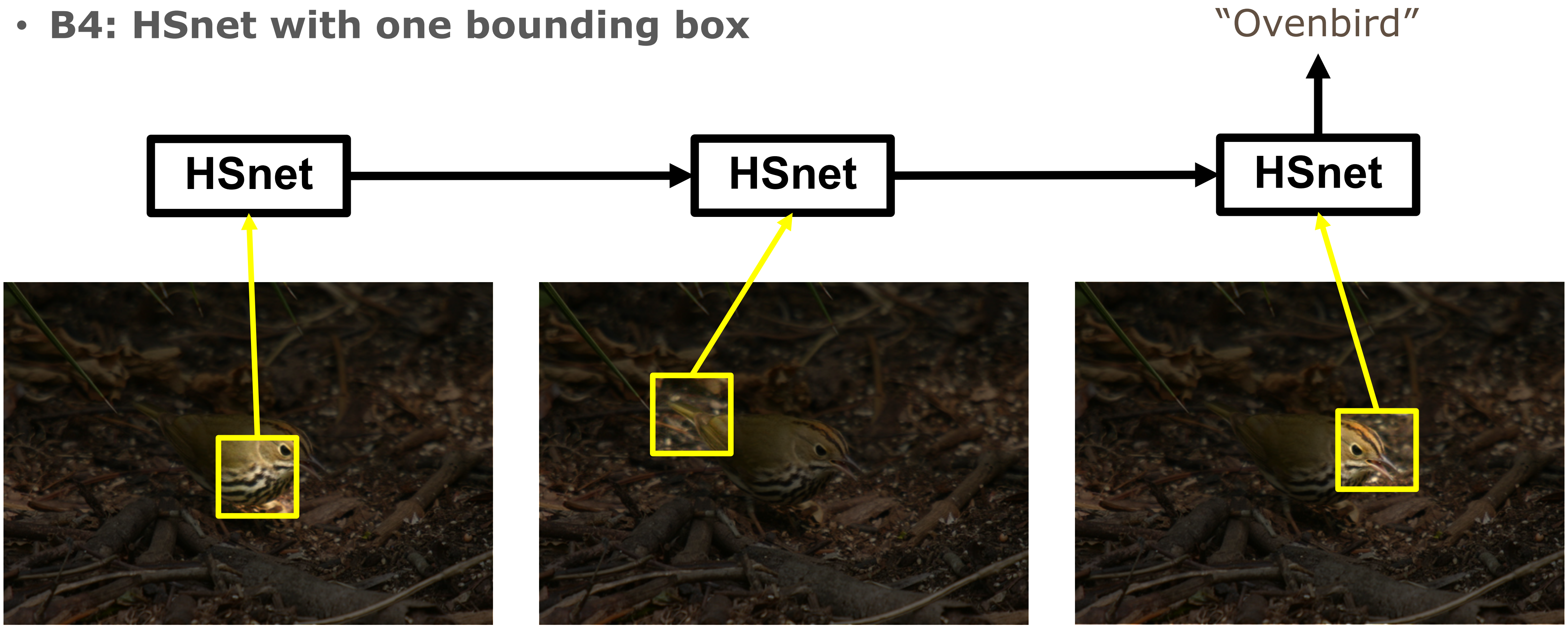
# Baselines

- B1: CNN (fine-tuned)
- B2: CNN with ground truth bounding boxes
- **B3: HSnet with one ground truth bounding box**
- B4: HSnet with one bounding box



# Baselines

- B1: CNN (fine-tuned)
- B2: CNN with ground truth bounding boxes
- B3: HSnet with one ground truth bounding box
- **B4: HSnet with one bounding box**



# Results: Caltech UCSD 2011 Birds

| Method | Annotations Used | Accuracy |
|---|---|---|
| Krause et al. 2015 | GT+BB | 82.8 |
| Jaderberg et al. 2015 | GT | 84.1 |
| Xu et al. 2015 | GT+BB+parts+web | 84.6 |
| Lin et al. 2015 | GT+BB | 85.1 |
| B1 | GT | 82.3 |
| B2 | GT+parts | 83.1 |
| B3 | GT+parts | 86.2 |
| B4 | GT+parts | 85.7 |
| **HSnet** | **GT+parts** | **87.5** |


[1] Krause et al. Fine-grained recognition without part annotations. CVPR, 2015.
[2] Jaderberg et al. Spatial transformer networks. NIPS, 2015.
[3] Xu et al. Augmenting strong supervision using web data for fine-grained categorization. CVPR, 2015.
[4] Lin et al. Bilinear cnn models for fine-grained visual recognition. ICCV, 2015.

# Results: Cars 196

| Method | Annotations Used | Accuracy |
|---|---|---|
| Deng et al. 2013 | GT+BB | 63.6 |
| Krause et al. 2013 | GT+BB | 67.6 |
| Krause et al. 2014 | GT+BB | 73.9 |
| Lin et al. 2015 | GT | 91.3 |
| Krause et al. 2015 | GT+BB | 92.6 |
| B1 | GT | 88.5 |
| B4 | GT | 92.2 |
| **HSnet** | **GT** | **93.9** |

[1] Deng et al. Fine-grained crowdsourcing for fine-grained recognition. CVPR, 2013.
[2] Krause et al. 3d object representations for fine-grained categorization. ICCV Workshop, 2013.
[3] Krause et al. Learning features and parts for fine-grained recognition. ICPR, 2014.
[4] Lin et al. Bilinear cnn models for fine-grained visual recognition. ICCV, 2015.
[5] Krause et al. Fine-grained recognition without part annotations. CVPR, 2015.

# Insights

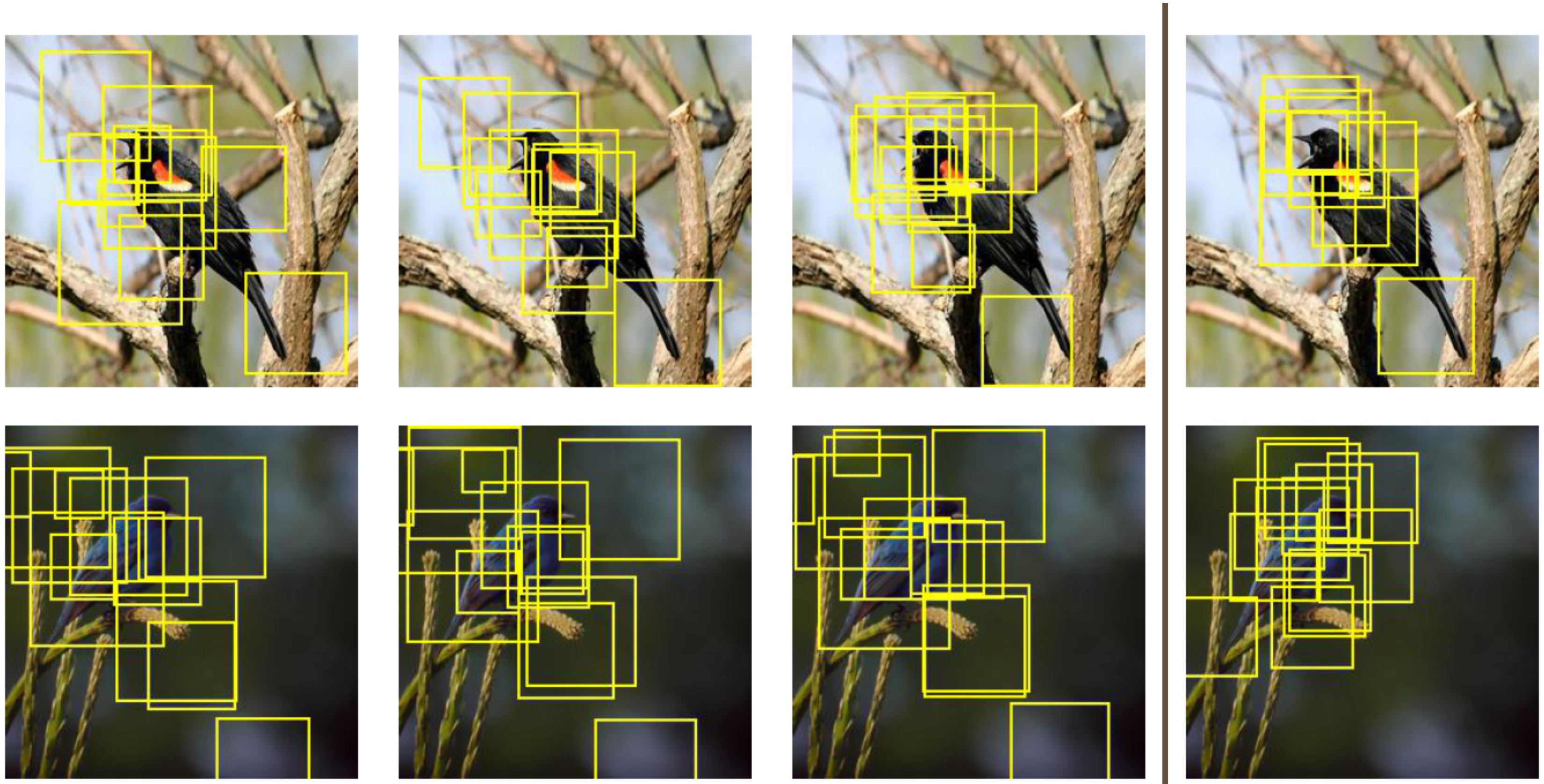
**Why is LSTM needed?**

- Baselines demonstrate that sequential reasoning (B3-B4) improves over one shot reasoning (B1-B2)

**Why DPP?**

- Regularization when no groundtruth part locations are provided
- Encourages learning diverse proposals rather than learning to single into one part

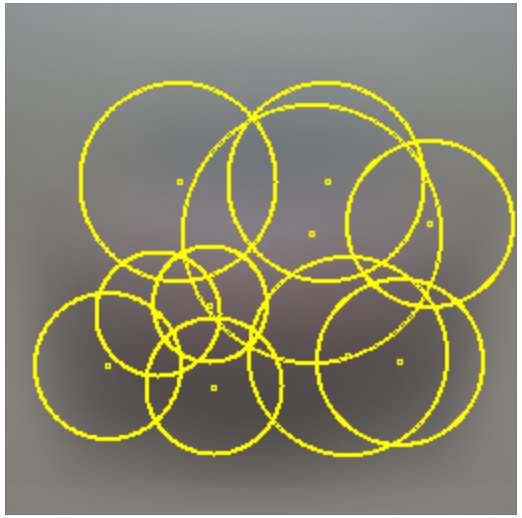
# Qualitative Results



$\tau = 5$ | $\tau = 10$ | $\tau = 15$ | Ground Truth

# Qualitative Results

Average Image of Cars



Clusters of Parts

# Summary

- Sequential search for informative image parts improves recognition
- DPP regularization works well when no parts annotations are provided
- Unlike most object recognition, fine-grained recognition benefits from focusing on parts

# Questions?

Fine-Grained Recognition as HSnet Search for Informative Image Parts

Michael Lam, Behrooz Mahasseni and Sinisa Todorovic
Oregon State University

CVPR 2017