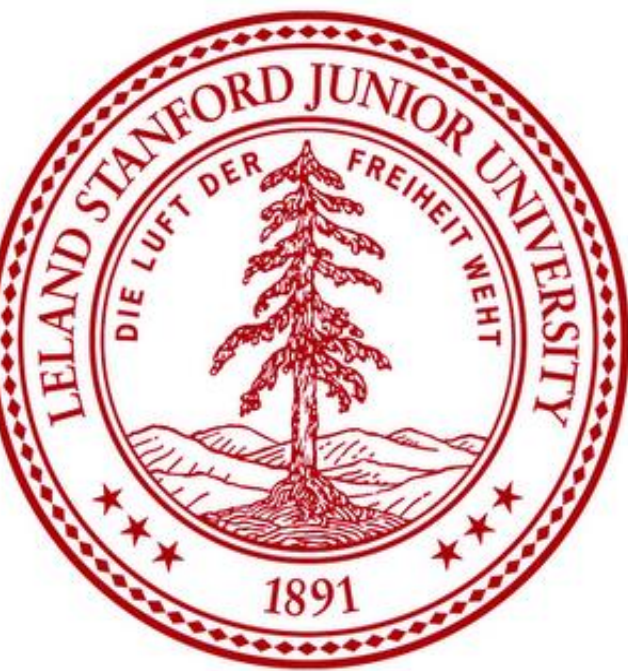# Co-localization in Real-World Images

Kevin Tang[1], Armand Joulin[1], Li-Jia Li[2], Li Fei-Fei[1]

[1]Computer Science Department, Stanford University
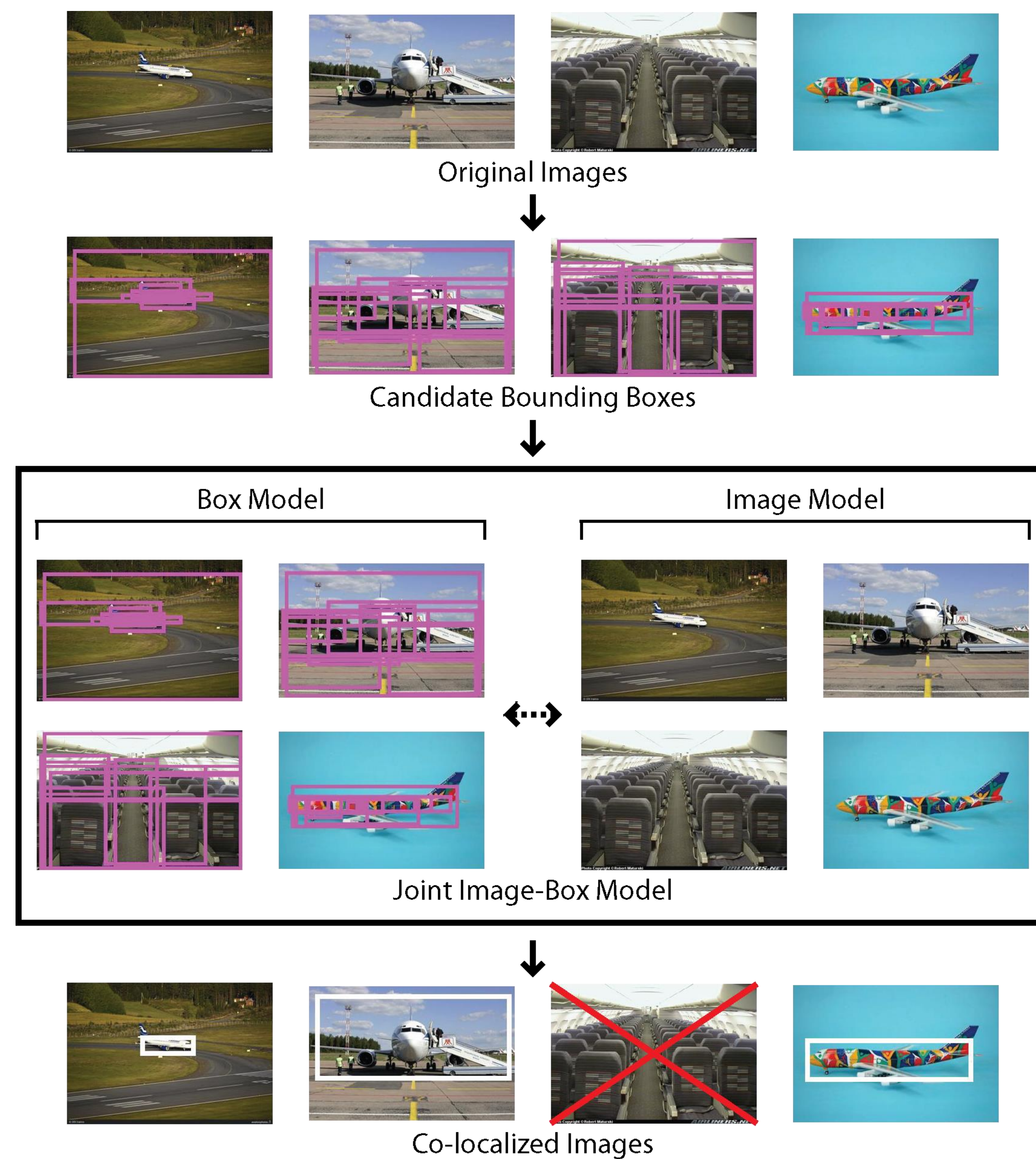
[2]Yahoo! Research

Introduction

- Goal
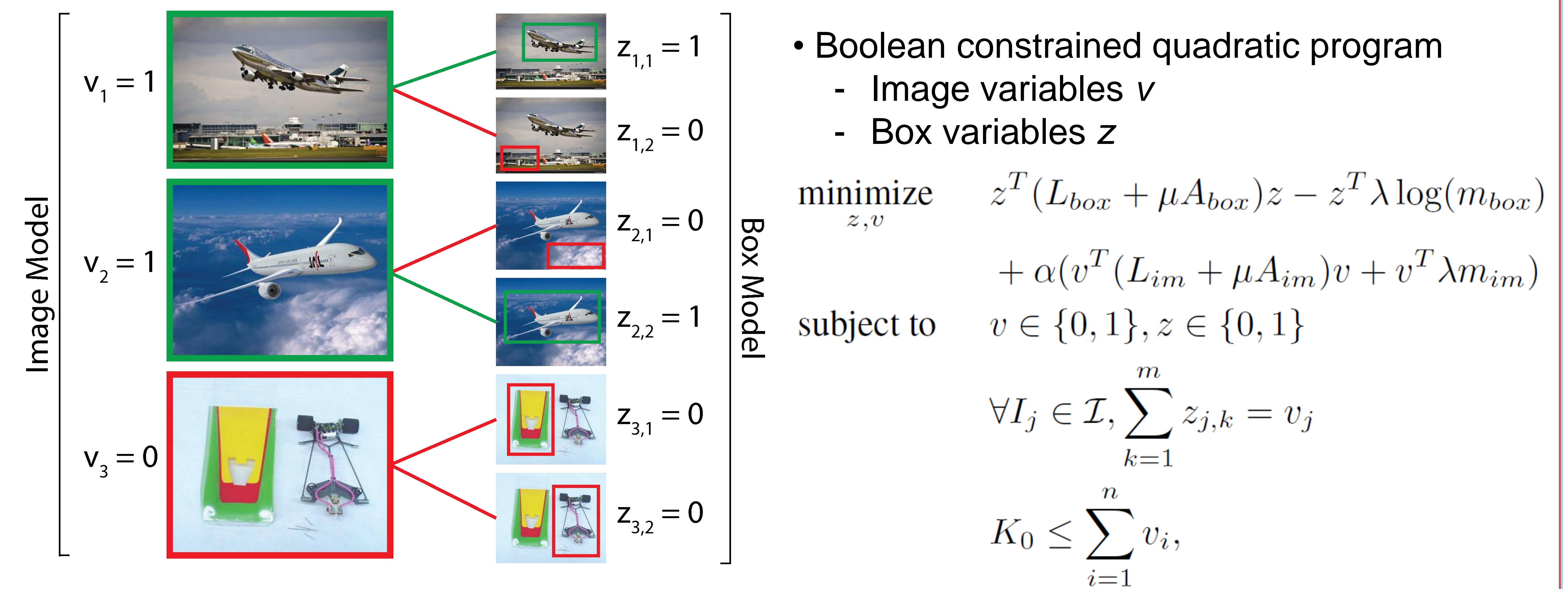 - Simultaneously localize (with boxes) objects of the same class across a set of distinct images

- Real-World Images
  - Intra-class variation of objects
  - Inter-class diversity of objects
  - Annotation noise in image tags/labels

Our Approach

Original Images

Candidate Bounding Boxes

Box Model / Image Model

Joint Image-Box Model

Co-localized Images

PASCAL VOC 2007

Method	Average CorLoc
Russell et al. [25]	22
Chum and Zisserman [7]	33
Deselaers et al. [9]	37
Our Method (prior)	13
Our Method (prior+similarity)	31
Our Method (full)	**39**

- PASCAL07-6x2 subset [9]
  - aeroplane, bicycle, boat, bus, horse, motorbike
  - 12 class/viewpoint combinations, 463 images total

Object Discovery Dataset

- 100 image subset [24]
  - airplane, car, horse
  - Noisy images that may not contain object

- CorLoc measurement [9]
  - Percentage of images correctly localized according to PASCAL criterion

Method	Average CorLoc
Kim et al. [17]	12.69
Joulin et al. [15]	51.35
Joulin et al. [16]	58.02
Rubinstein et al. [24]	75.16
Our Method	**76.58**

Precision-recall curves for identifying noisy images

(a)

(b)

Joint Image-Box Model

- Boolean constrained quadratic program
  - Image variables v
  - Box variables z

$$\underset{z,v}{\text{minimize}} \quad z^T(L_{box} + \mu A_{box})z - z^T \lambda \log(m_{box})$$
$$+ \alpha(v^T(L_{im} + \mu A_{im})v + v^T \lambda m_{im})$$

$$\text{subject to} \quad v \in \{0,1\}, z \in \{0,1\}$$

$$\forall I_j \in \mathcal{I}, \sum_{k=1}^{m} z_{j,k} = v_j$$

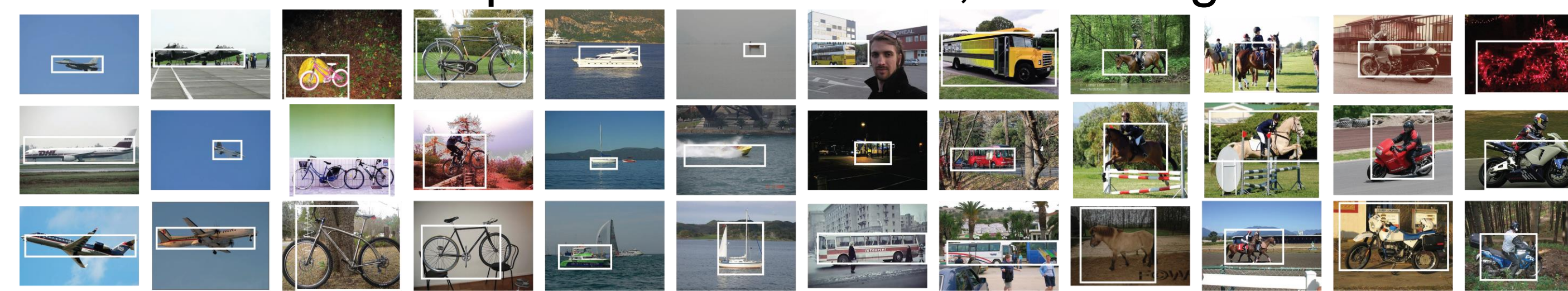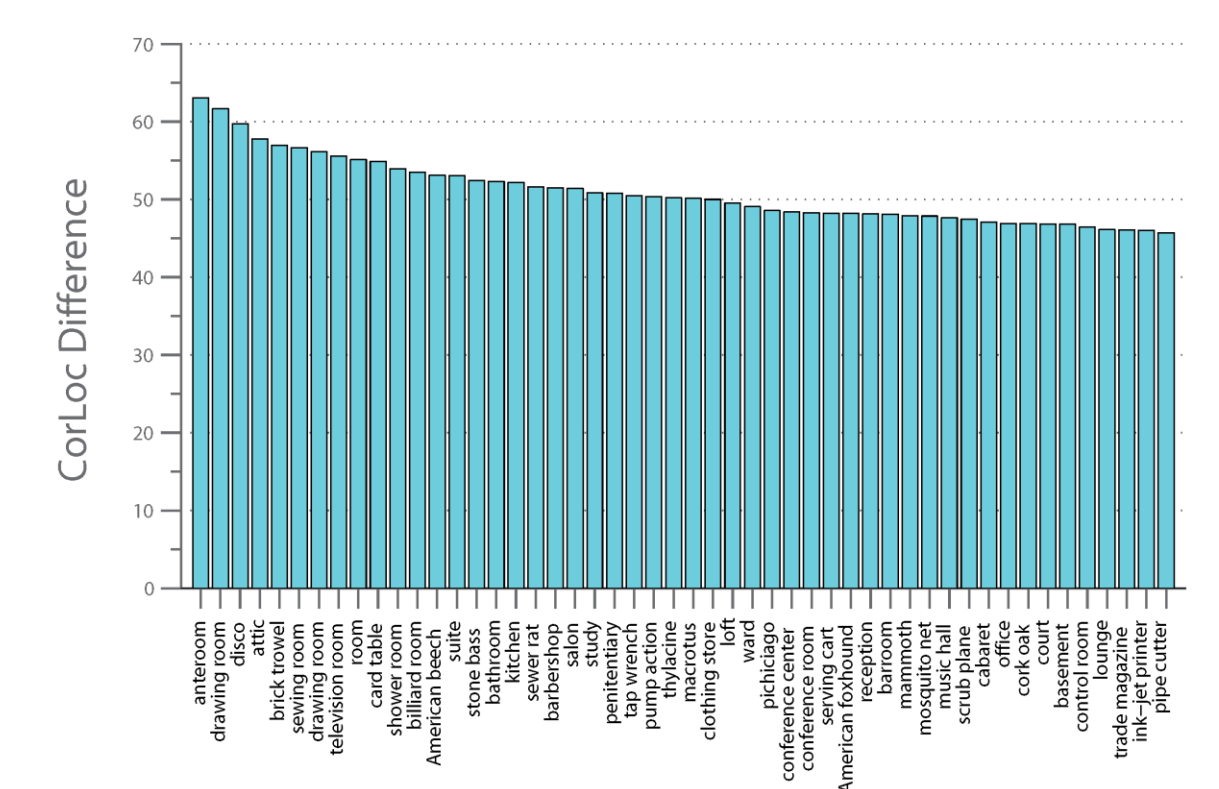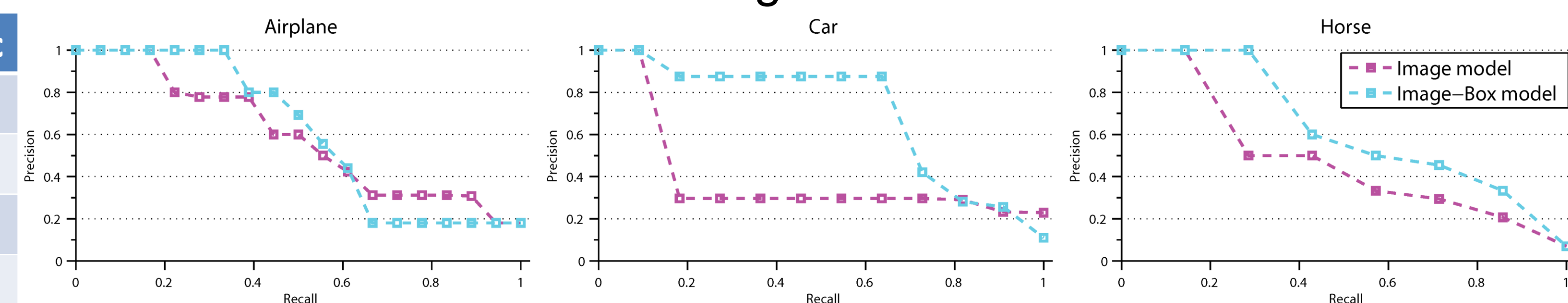$$K_0 \leq \sum_{i=1}^{n} v_i,$$

- **Prior** (linear term): m vector
  - Image prior: chi-squared distance to average image feature in set
  - Box prior: average saliency weighted by size of box

- **Similarity** (quadratic term): L matrix
  - Encourage images/boxes with similar appearances to have the same label
  - Normalized Laplacian matrix computed from chi-squared similarity matrix

- **Discriminability** (quadratic term): A matrix
  - Encourage selecting images/boxes that are discriminatively separable from others
  - Closed-form solution to ridge regression objective function

ImageNet

- Co-localize all images with ground-truth bounding box annotations
  - 3,624 classes / 939,542 images

Method	Average CorLoc
Top objectness box [1]	37.42
Our Method	**53.20**

Example results: our method (white), ground-truth (green)
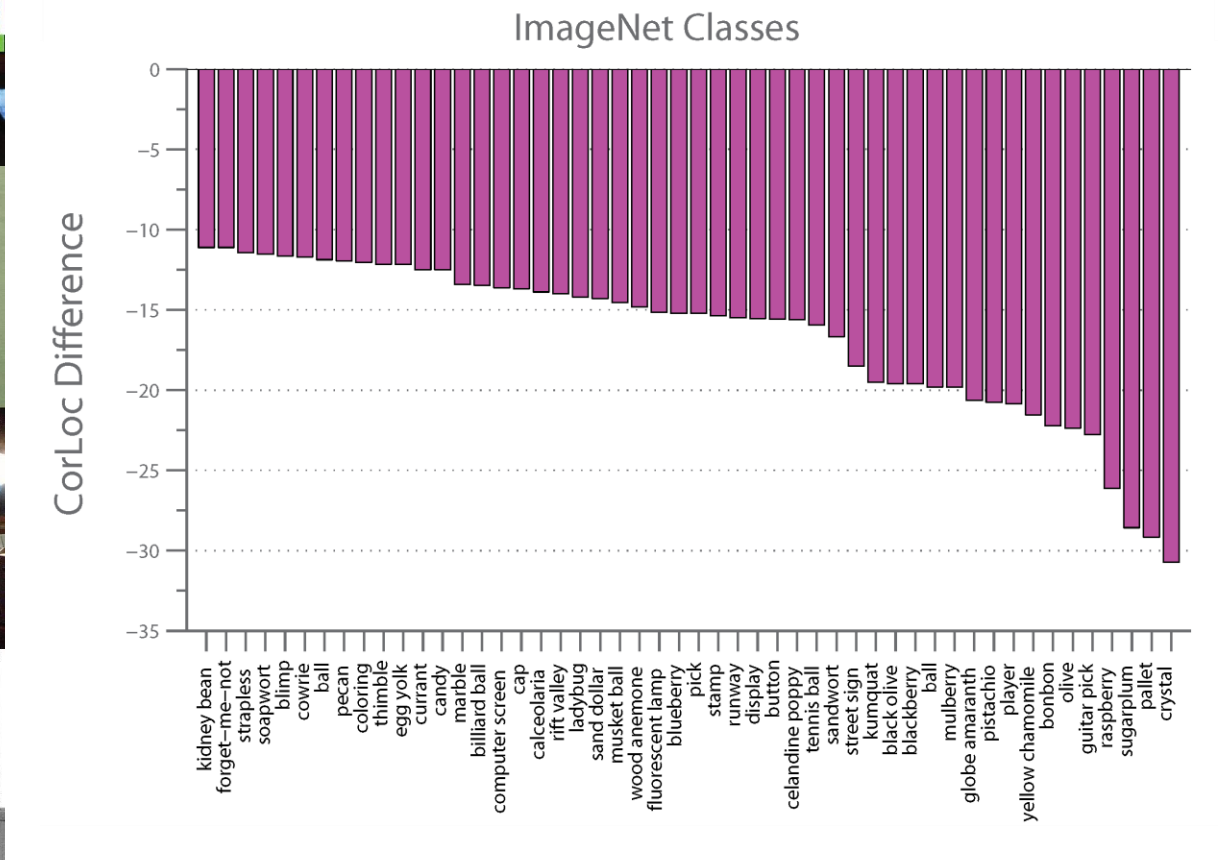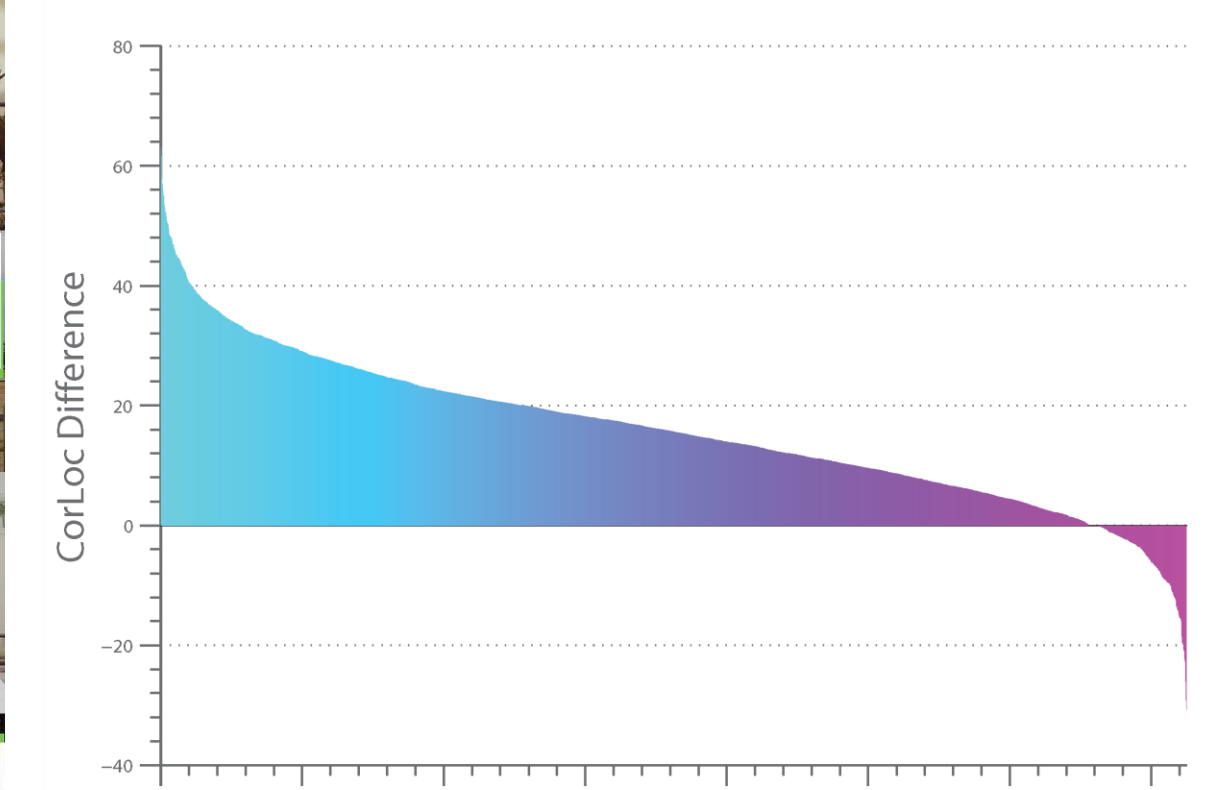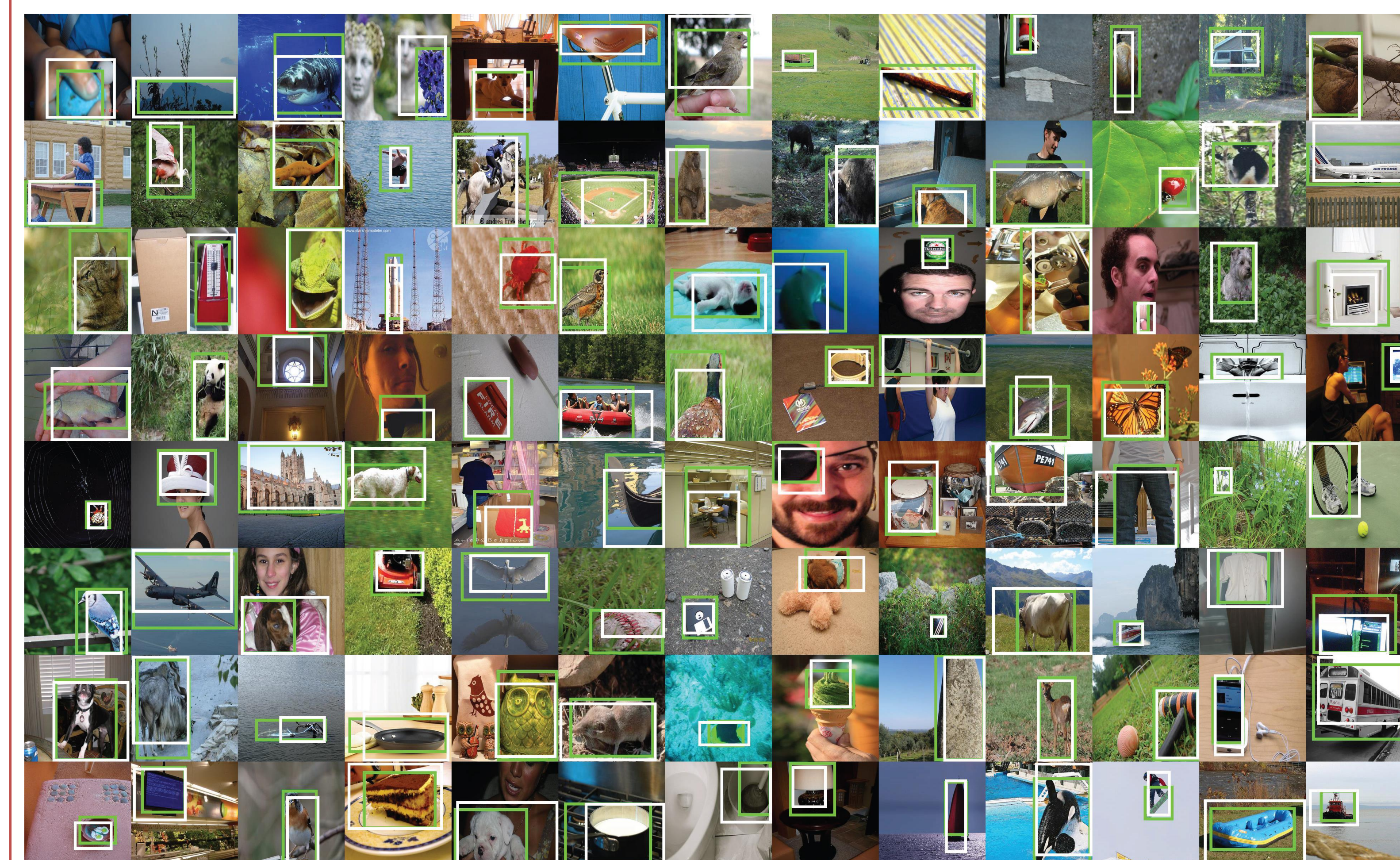
Distribution of best/worst CorLoc differences