How to Collect Segmentations for Biomedical Images?
A Benchmark Evaluating the Performance of Experts, Crowdsourced Non-Experts, and Algorithms

Danna Gurari[1], Diane Theriault[1], Mehrnoosh Sameki[1], Brett Isenberg[2], Tuan A. Pham[2], Alberto Purwada[2], Patricia Solski[2], Matthew Walker[3], Chentian Zhang[2],
Joyce Y. Wong[2], Margrit Betke[1]; [1] *Department of Computer Science*; [2] *Department of Biomedical Engineering*; [3] *Department of Biology*

Motivation

Biology: Relationship between ***shape*** and ***function***?

Question: How to extract accurate object boundaries from images?

Key Challenges:

* Which annotation method?

* How to evaluate an annotation method? (no references for "images in the wild")

Image Library

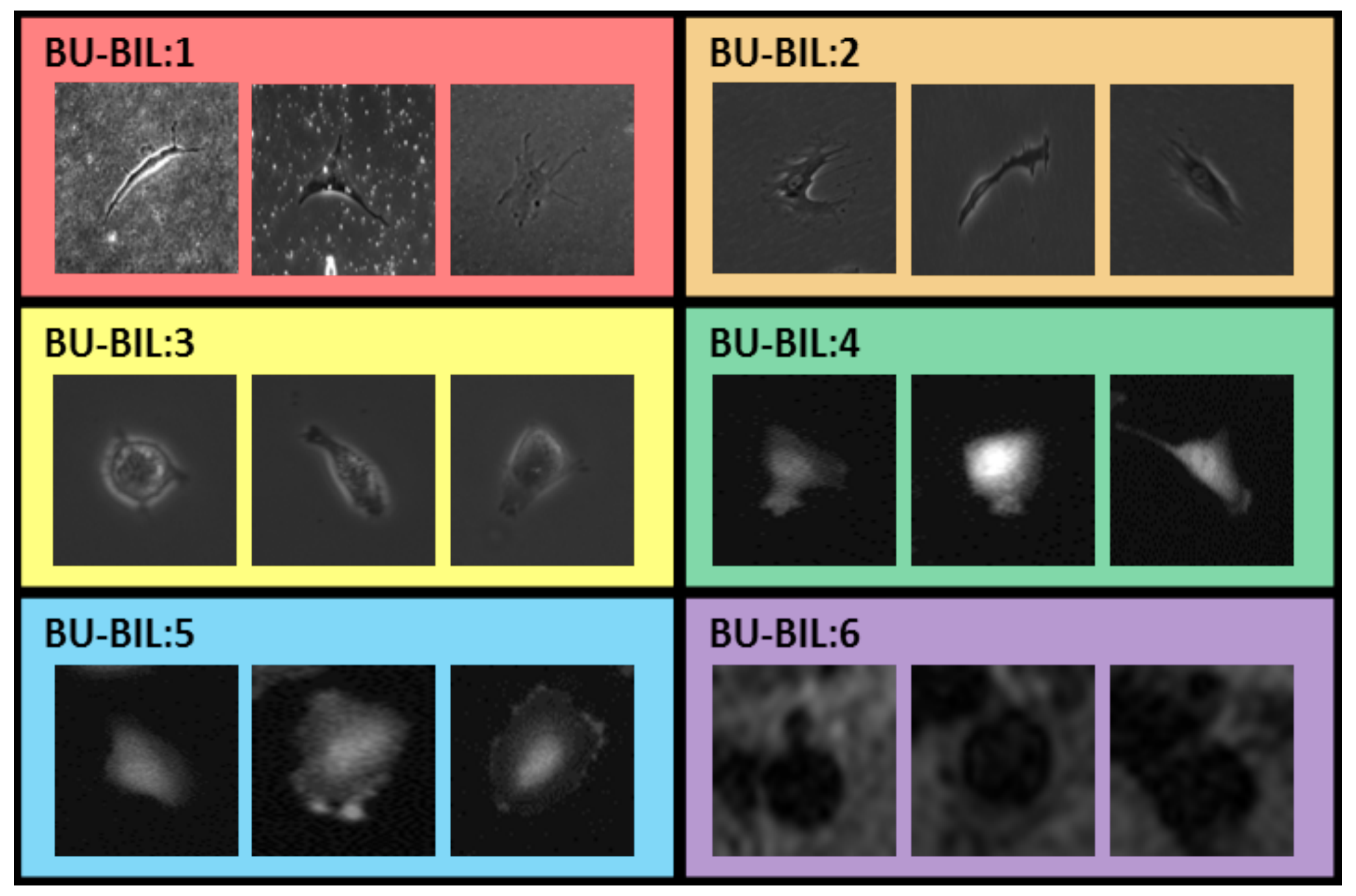


Table 1. Salient properties characterizing each dataset in the image library and the number of objects per dataset.

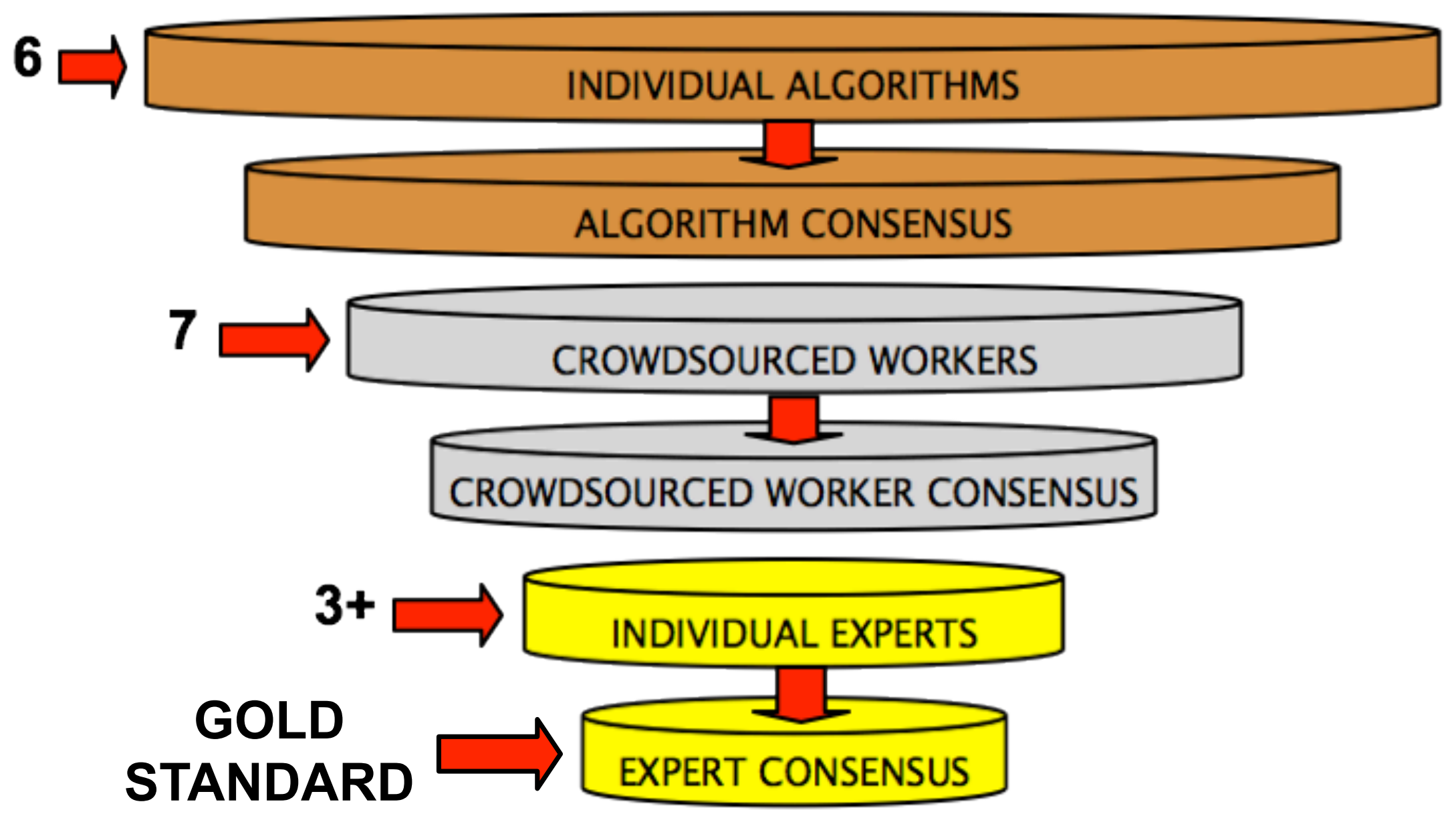
ID	Modality	Object Type	Mag.	Avg. Pixel Count	Avg. Circularity	Avg. Object Intensity	Avg. Bkgrnd Intensity	# Objs
1	Phase Contrast	Rat smooth muscle cells	40x	35,613	0.15	64	61	35
2	Phase Contrast	Rabbit smooth muscle cells	10x	10,963	0.29	52	50	69
3	Phase Contrast	Fibroblasts	10x	3,937	0.53	58	50	47
4	Fluorescence	Lu melanoma cells	10x	836	0.53	48	17	61
5	Fluorescence	WM993 melanoma cells	10x	1,119	0.71	54	19	58
6	MRI	Rabbit aorta	10x	216	0.94	25	42	35

3 image acquisition modalities

6 biological entities

305 Objects

Annotation Methods


6

INDIVIDUAL ALGORITHMS

ALGORITHM CONSENSUS

7

CROWDSOURCED WORKERS

CROWDSOURCED WORKER CONSENSUS

3+

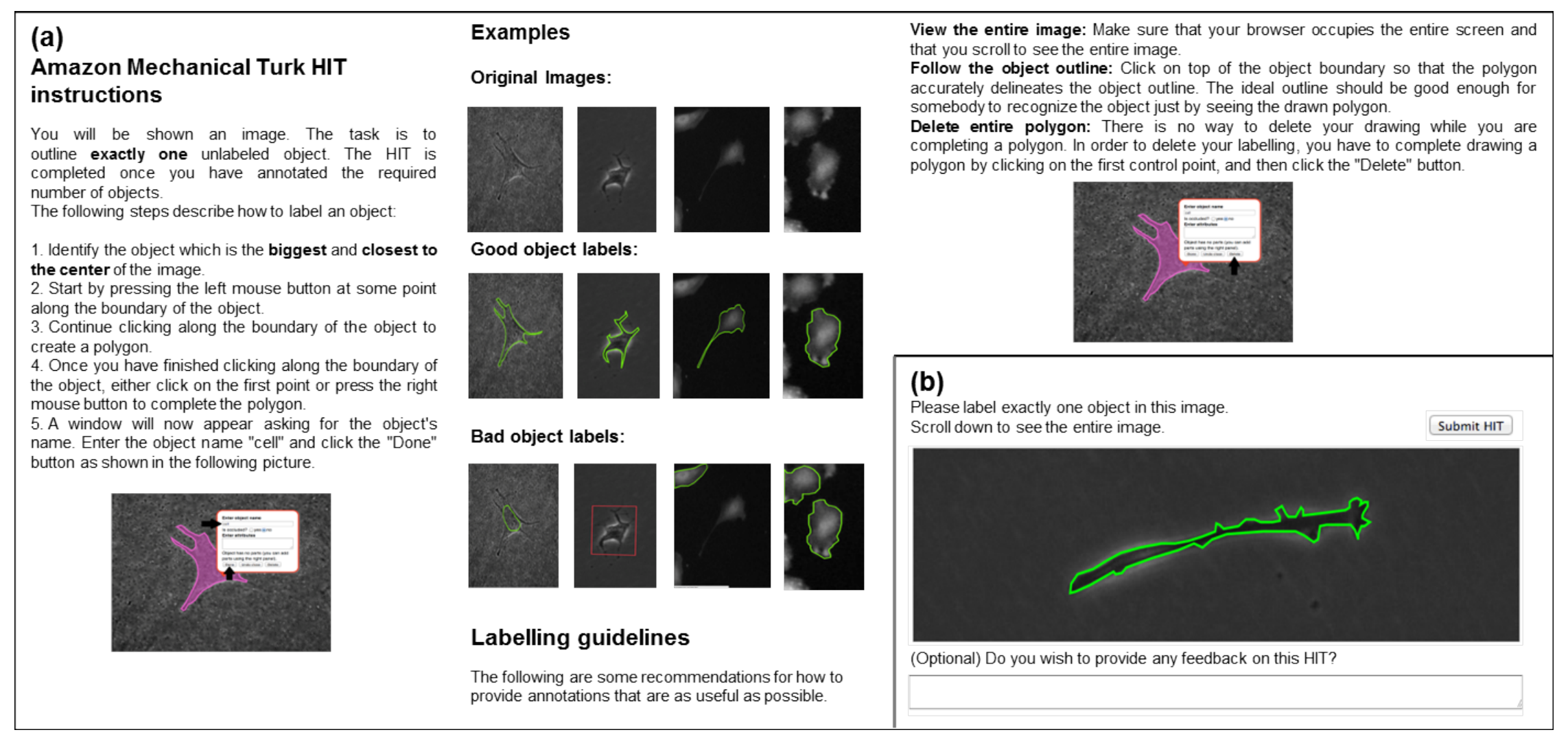
INDIVIDUAL EXPERTS

GOLD STANDARD

EXPERT CONSENSUS

Algorithms: Lankton level set, Shi level set, Chan Vese level set, seeded watershed, Hough Transform, Otsu thresholding

Crowdsourcing: LabelMe & Mechanical Turk



(a)
Amazon Mechanical Turk HIT instructions

You will be shown an image. The task is to outline **exactly one** unlabeled object. The HIT is completed once you have annotated the required number of objects.
The following steps describe how to label an object:

1. Identify the object which is the **biggest** and **closest to the center** of the image.
2. Start by pressing the left mouse button at some point along the boundary of the object.
3. Continue clicking along the boundary of the object to create a polygon.
4. Once you have finished clicking along the boundary of the object, either click on the first point or press the right mouse button to complete the polygon.
5. A window will now appear asking for the object's name. Enter the object name "cell" and click the "Done" button as shown in the following picture.

Examples

Original Images:

Good object labels:

Bad object labels:

Labelling guidelines

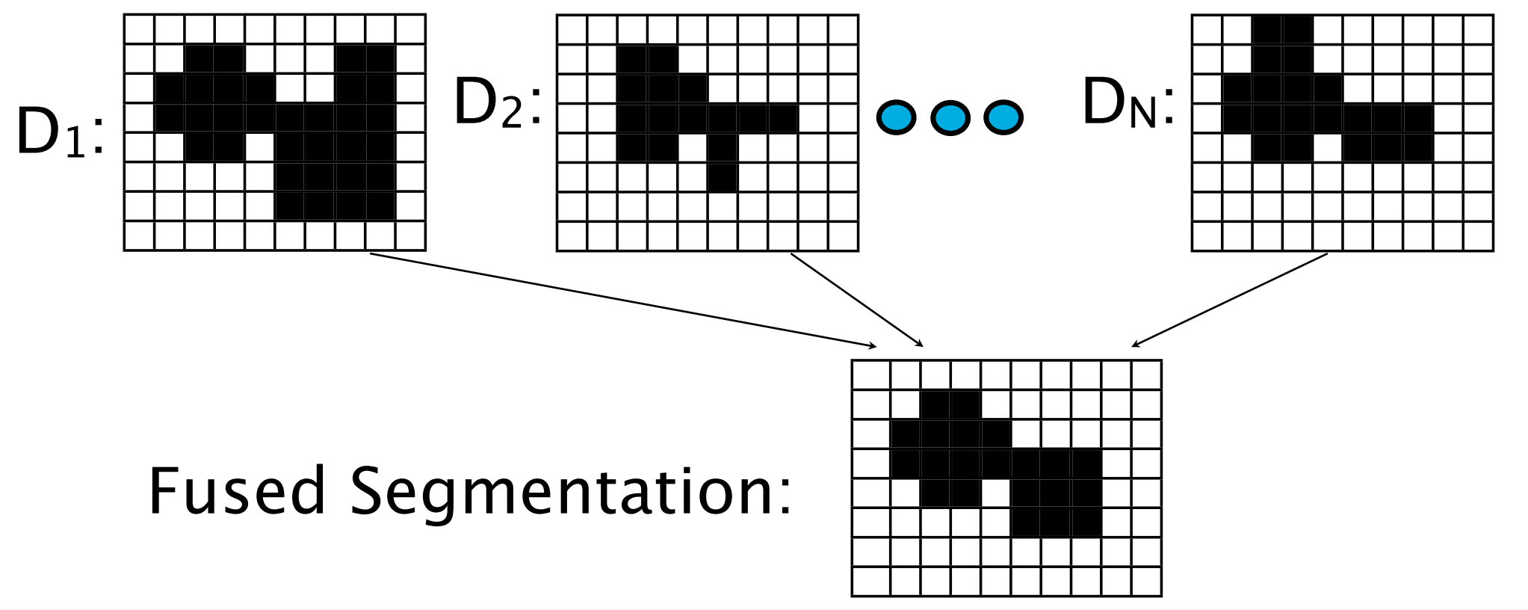
The following are some recommendations for how to provide annotations that are as useful as possible.

View the entire image: Make sure that your browser occupies the entire screen and that you scroll to see the entire image.
Follow the object outline: Click on top of the object boundary so that the polygon accurately delineates the object outline. The ideal outline should be good enough for somebody to recognize the object just by seeing the drawn polygon.
Delete entire polygon: There is no way to delete your drawing while you are completing a polygon. In order to delete your labelling, you have to complete drawing a polygon by clicking on the first control point, and then click the "Delete" button.

(b)
Please label exactly one object in this image.
Scroll down to see the entire image.

(Optional) Do you wish to provide any feedback on this HIT?

Consensus: Pixel Majority Vote

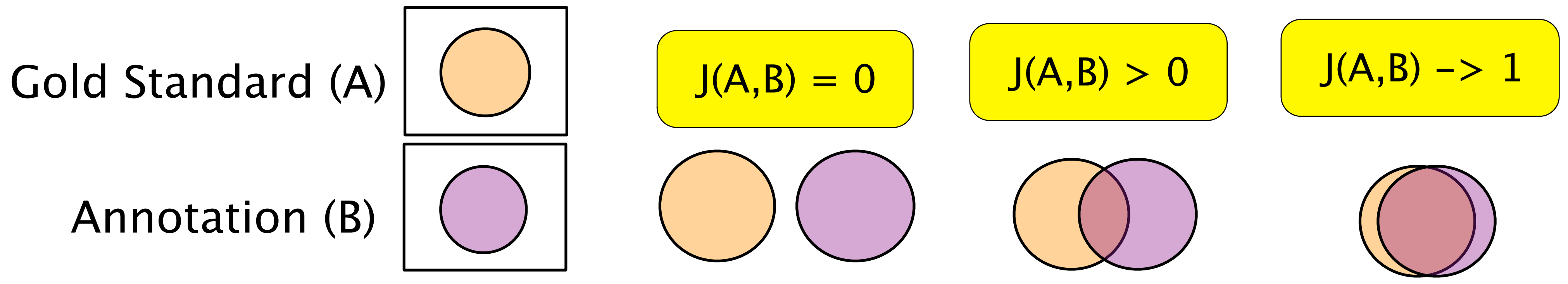


Acknowledgments

NSF Grants: IIS-1421943, IIS-0910908

Method Evaluation

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

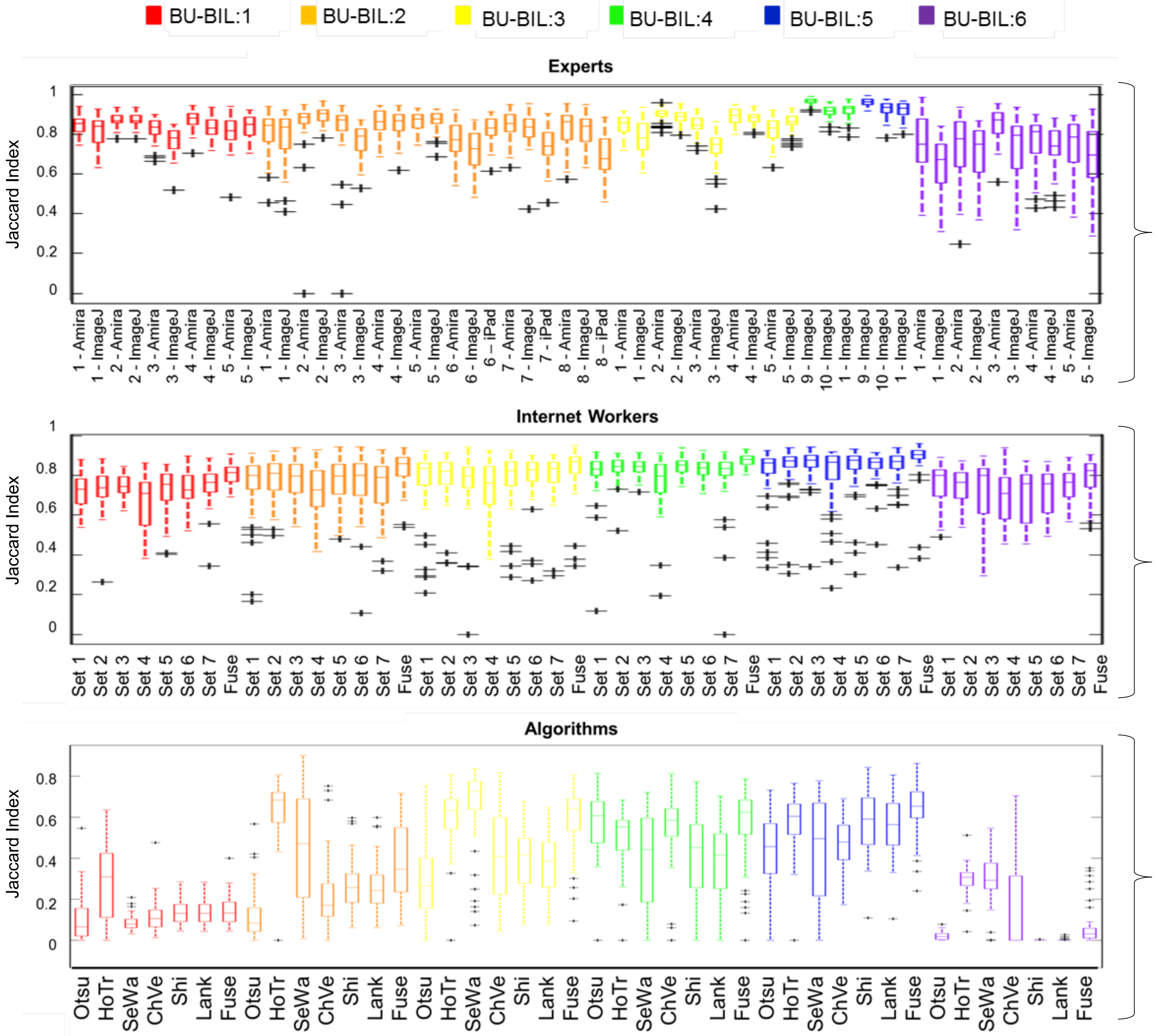

Gold Standard (A)

Annotation (B)

J(A,B) = 0

J(A,B) > 0

J(A,B) -> 1



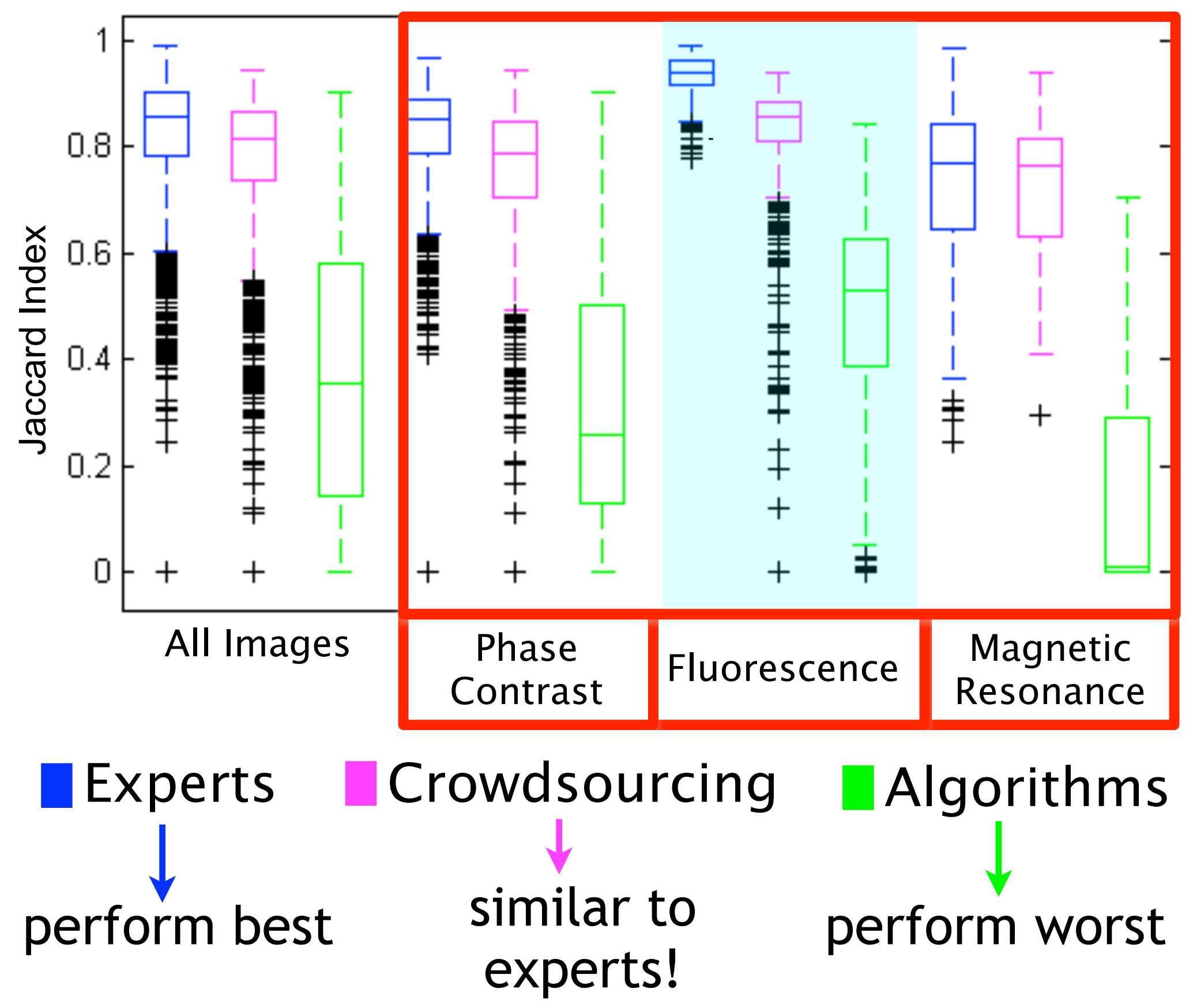
Expert performance differs, especially for different datasets

Crowd consensus exceeds performance of individual crowd workers

Algorithm performance varies widely, especially for different datasets

Method Comparison

Crowdsourcing consensus statistically similar to experts!



🟦 Experts — perform best

🟪 Crowdsourcing — similar to experts!

🟩 Algorithms — perform worst