Learning object categories for efficient bottom-up recognition

Daniel Kersten
Psychology Department, University of Minnesota

kersten.org

Supported by ONR N 00014-07-1-0937

Challenge of complexity in natural image input

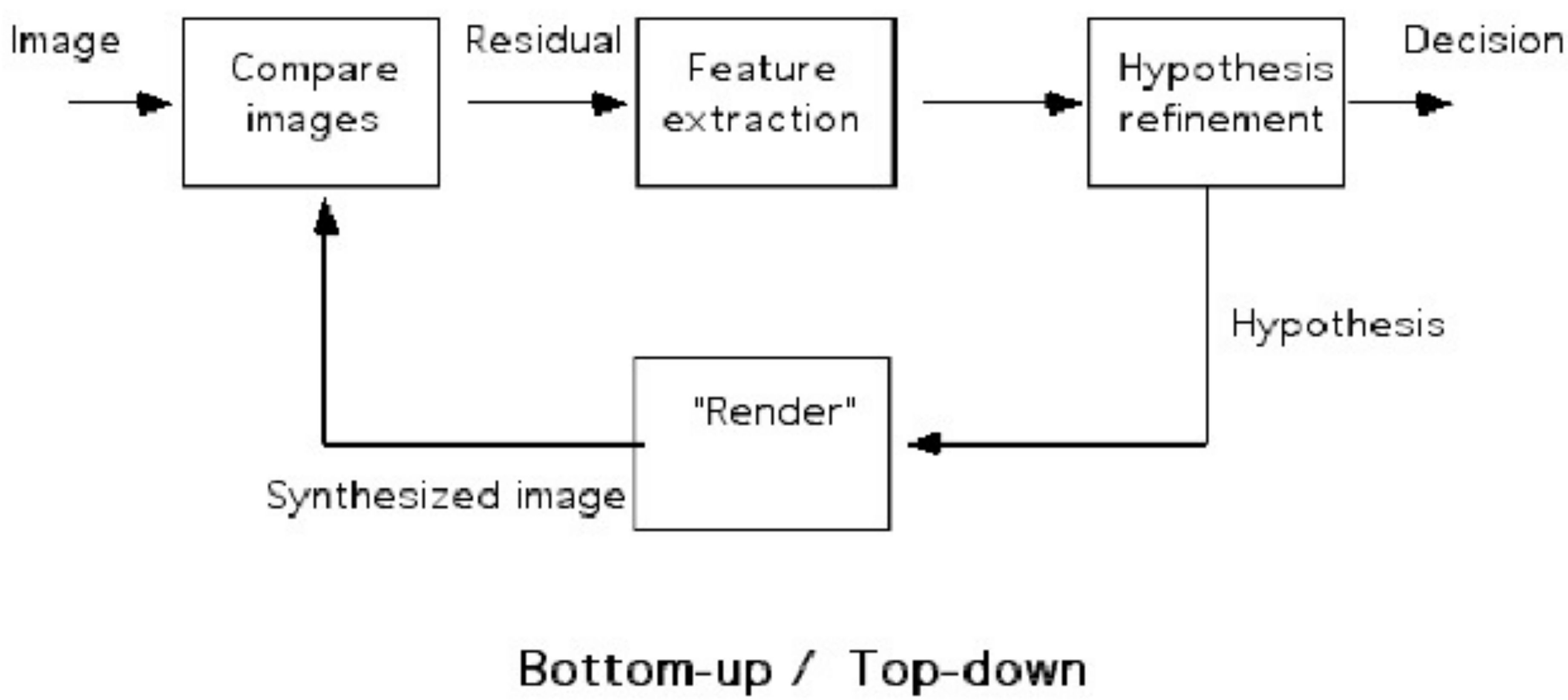
- Enormous range of variability in the images for a given object category, eg. "foxes"
- Enormous objective uncertainty regarding image features present for any given exemplar

How to learn to be maximally effective across a broad range of tasks?

- Need generative "world model" that can account for previously unexperienced combinations of objects, background, lighting, pose, ...
- Need efficient selection of critical diagnostic features to index object classes that will *generalize across all within-class instances*
- *Learning object categories*
- The challenge of learning from a small number of examples

Mechanisms for flexible recognition

- Generative mechanisms: "Analysis by Synthesis"



Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? Trends Cogn Sci, 10(7), 301-308.

For recognition, analysis by synthesis useful when:

- Segmentation in cluttered scenes
- Transformations that are computationally difficult to do bottom-up, e.g.
  - orientation in 3D depth
  - articulations, e.g. scissors
  - occlusion
- Competing/interacting object property/scene hypotheses

Computational Example

Three models: text, faces, texture

Input

Tu, Z., Chen, X., Yuille, A., & Zhu, S. (2005). Image Parsing: Unifying Segmentation, Detection and Recognition. IJCV, 63(2).

Computational Example

Three models: text, faces, texture

Input

Tu, Z., Chen, X., Yuille, A., & Zhu, S. (2005). Image Parsing: Unifying Segmentation, Detection and Recognition. IJCV, 63(2).

Computational Example

Three models: text, faces, texture

Input

Tu, Z., Chen, X., Yuille, A., & Zhu, S. (2005). Image Parsing: Unifying Segmentation, Detection and Recognition. IJCV, 63(2).

Computational Example

Three models: text, faces, texture

Input

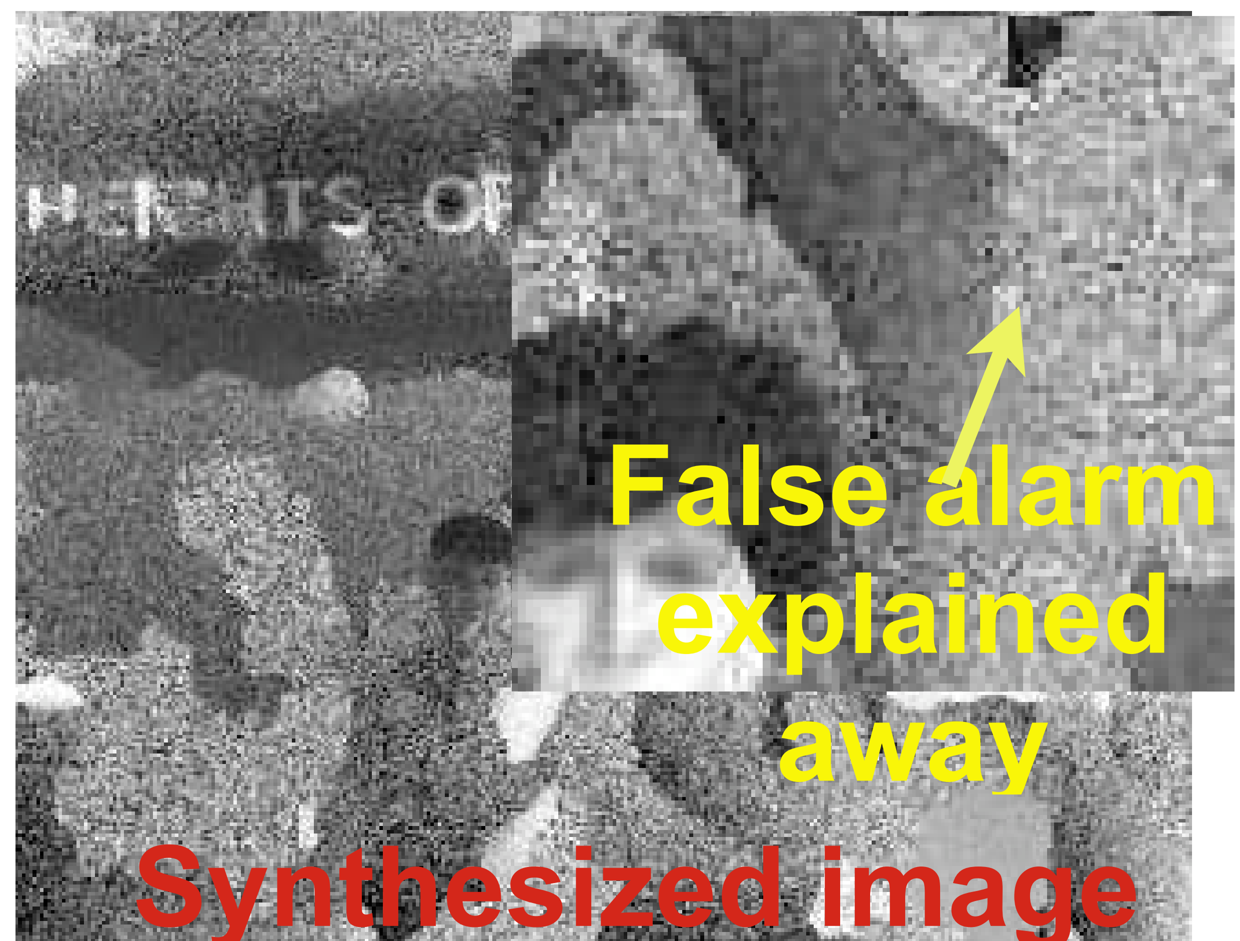
Tu, Z., Chen, X., Yuille, A., & Zhu, S. (2005). Image Parsing: Unifying Segmentation, Detection and Recognition. IJCV, 63(2).

Computational Example

Three models: text, faces, texture

Input

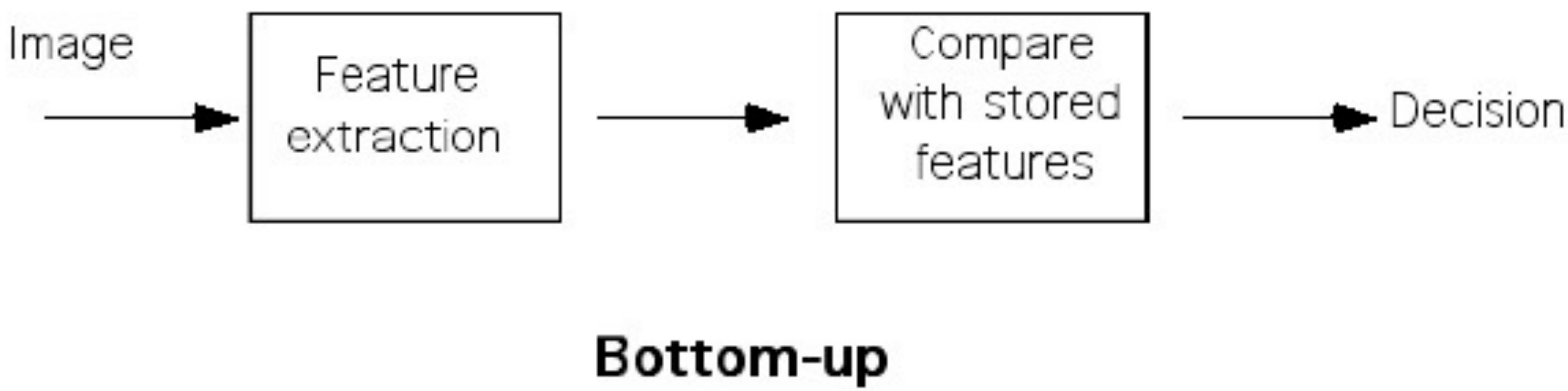
Tu, Z., Chen, X., Yuille, A., & Zhu, S. (2005). Image Parsing: Unifying Segmentation, Detection and Recognition. IJCV, 63(2).



Strategies

- Generative mechanisms
  - provide flexibility
- ...BUT computational/behavioral speed and accuracy requires effective diagnostic features to deal with the enormous with-class variation within a pattern/object category

“Discriminative models”



Need to learn features (“index features”) to support reliable if not perfect first, bottom-up pass

How to learn features to support a variety of actions, not just decisions about labels

- Size perception, e.g. for interception
- Material, e.g. for driving
- ...
- Object categorization
  - Do discriminative features learned in one task transfer to another?

How to learn features to support a variety of actions, not just decisions about labels

- Size perception, e.g. for interception
- Material, e.g. for driving
- ...
- Object categorization
  - Do discriminative features learned in one task transfer to another?

Computational example: Learning informative features for a task

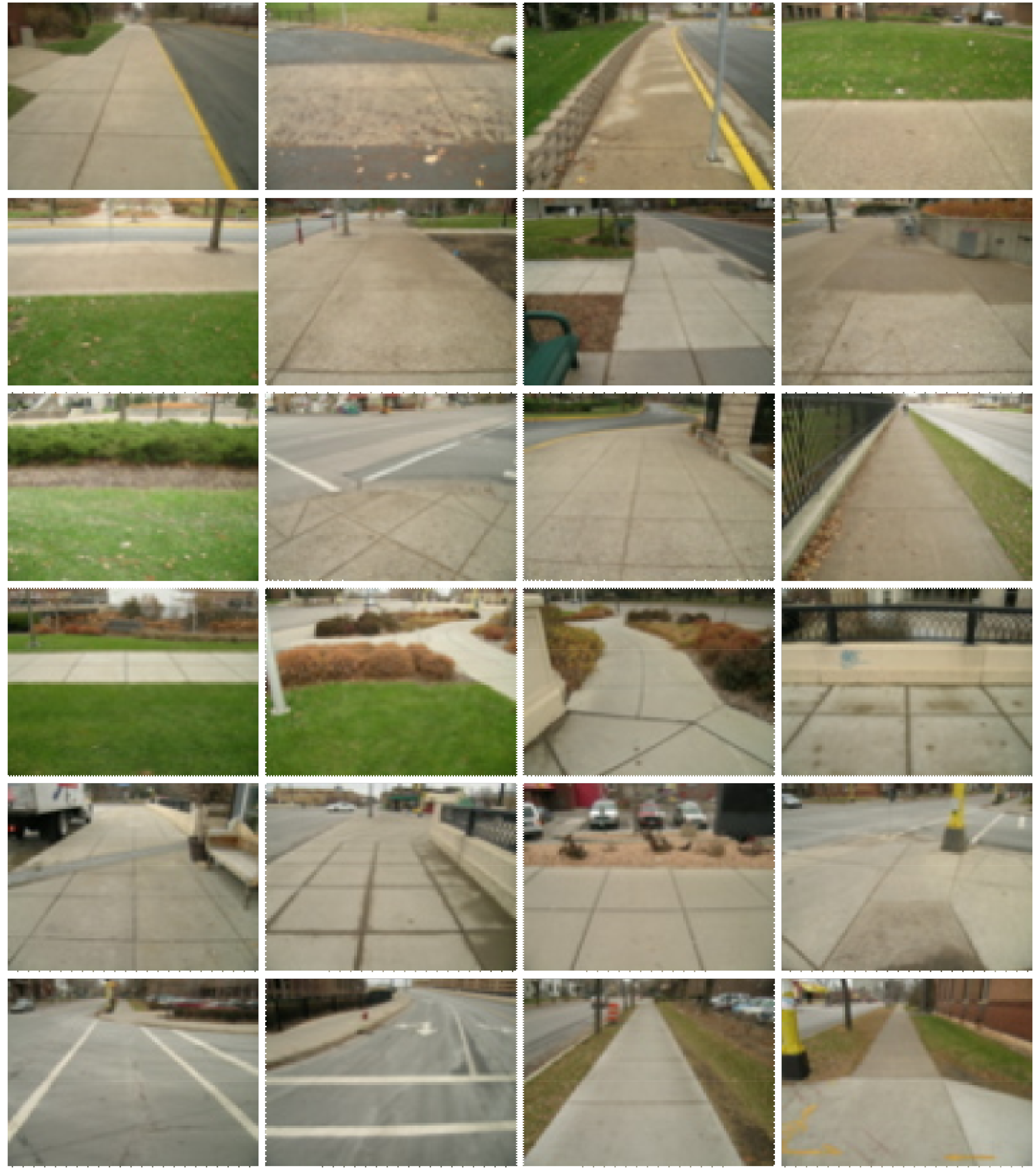
What do these scenes have in common?

With Evgeniy Bart

“Up” curbs-- that require a step up

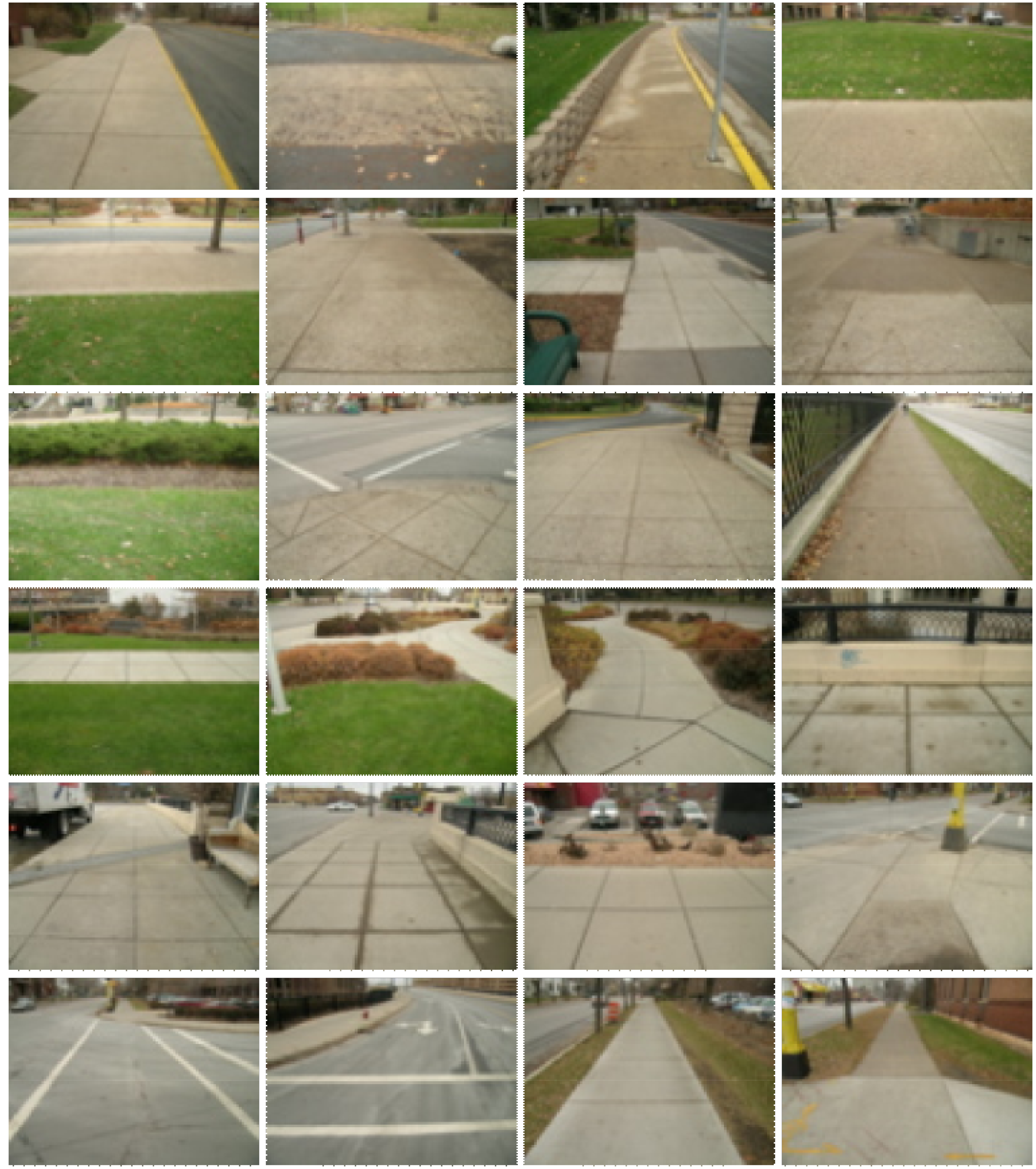
“Up” curbs-- that require a step up

Distinguish from Non- “up curbs”



Distinguish from Non- “up curbs”

...that do not require a step



Distinguish from Non- “up curbs”

...that do not require a step

Selecting diagnostic features

$$I(C;F) = H(C) - H(C|F)$$

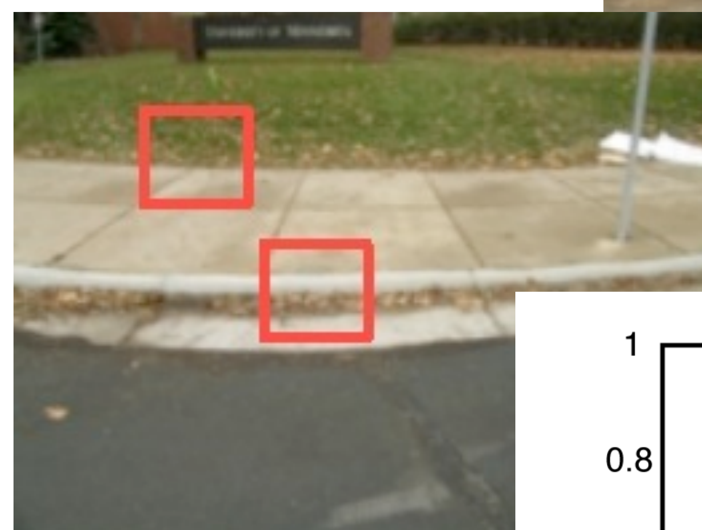
$$\begin{aligned} F_1 &= \arg\max_F I(C;F); \\ F_{k+1} &= \arg\max_F \min_i I(C;F|F_i) \end{aligned}$$

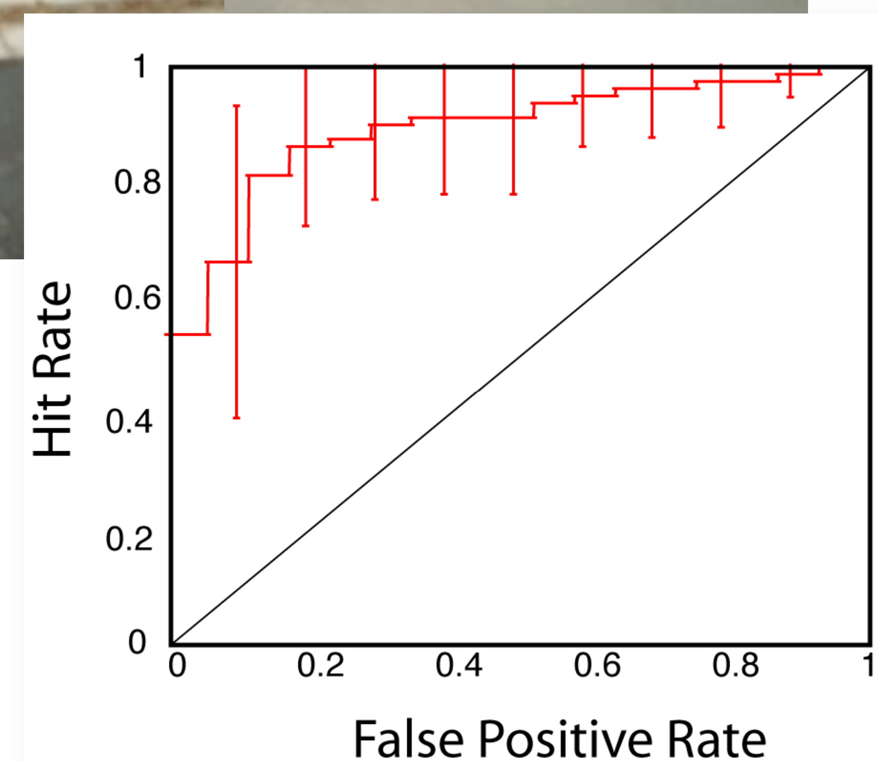
Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. Nat Neurosci, 5(7), 682-687.

Learning based on informative fragments for the task

- Find fragments that maximize mutual information (Ullman et al., 2002; Bart et al, 2004)

- Detect "up curbs" from an approach angle that requires a step





With Evgeniy Bart

Learning object categories

Do image features (fragments) that maximize mutual information predict the features that human observers learn to use?

Need novel object classes with small within-class variation and slightly larger between-class variation

Virtual phylogenesis of digital embryos

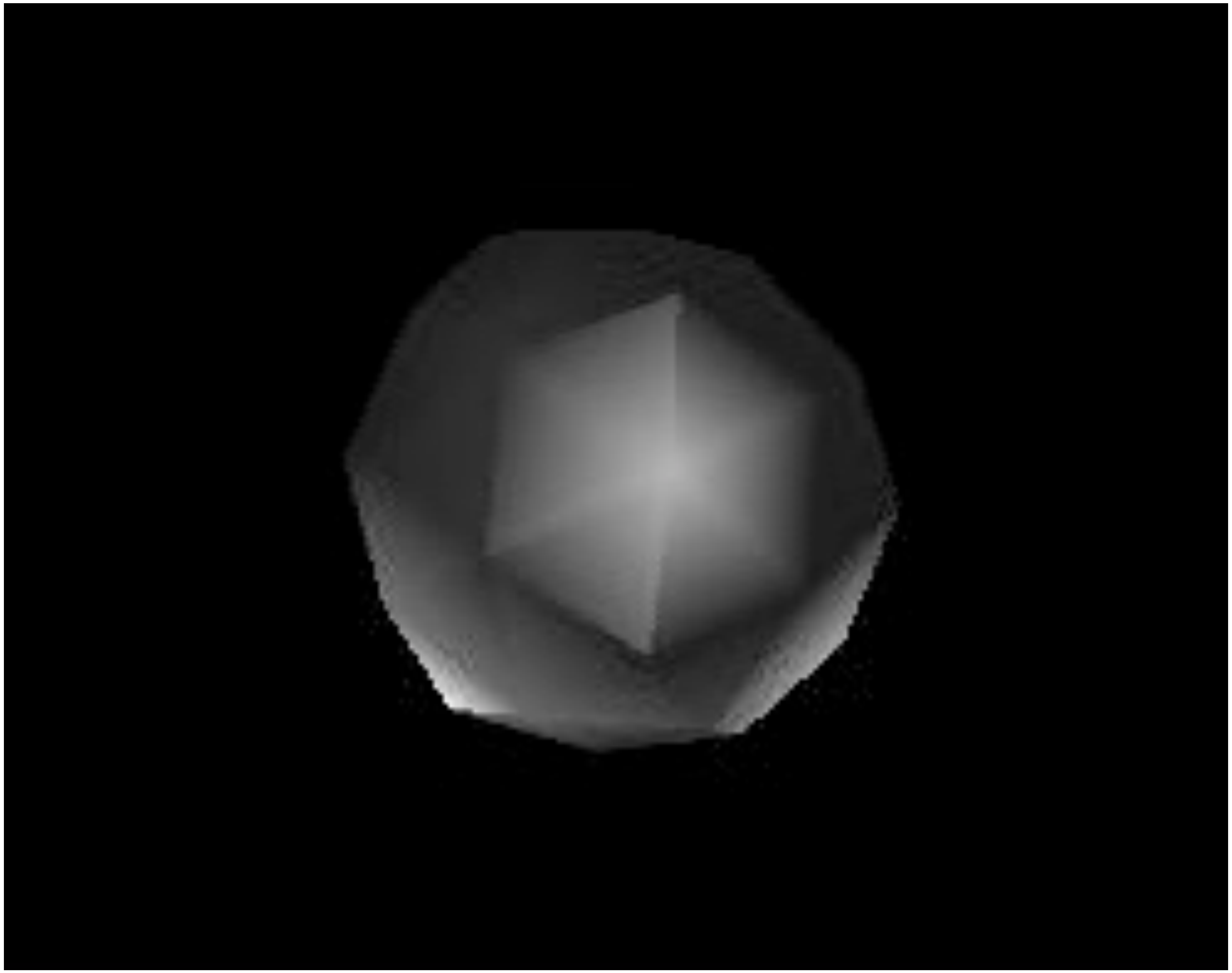
Hegde, J., Bart, E., & Kersten, D. (2008). Fragment-Based Learning of Visual Object Categories. Curr Biol. 18, 597-601

Digital embryo growth

Prof. Mark Brady

http://www.psych.ndsu.nodak.edu/brady/downloads.html

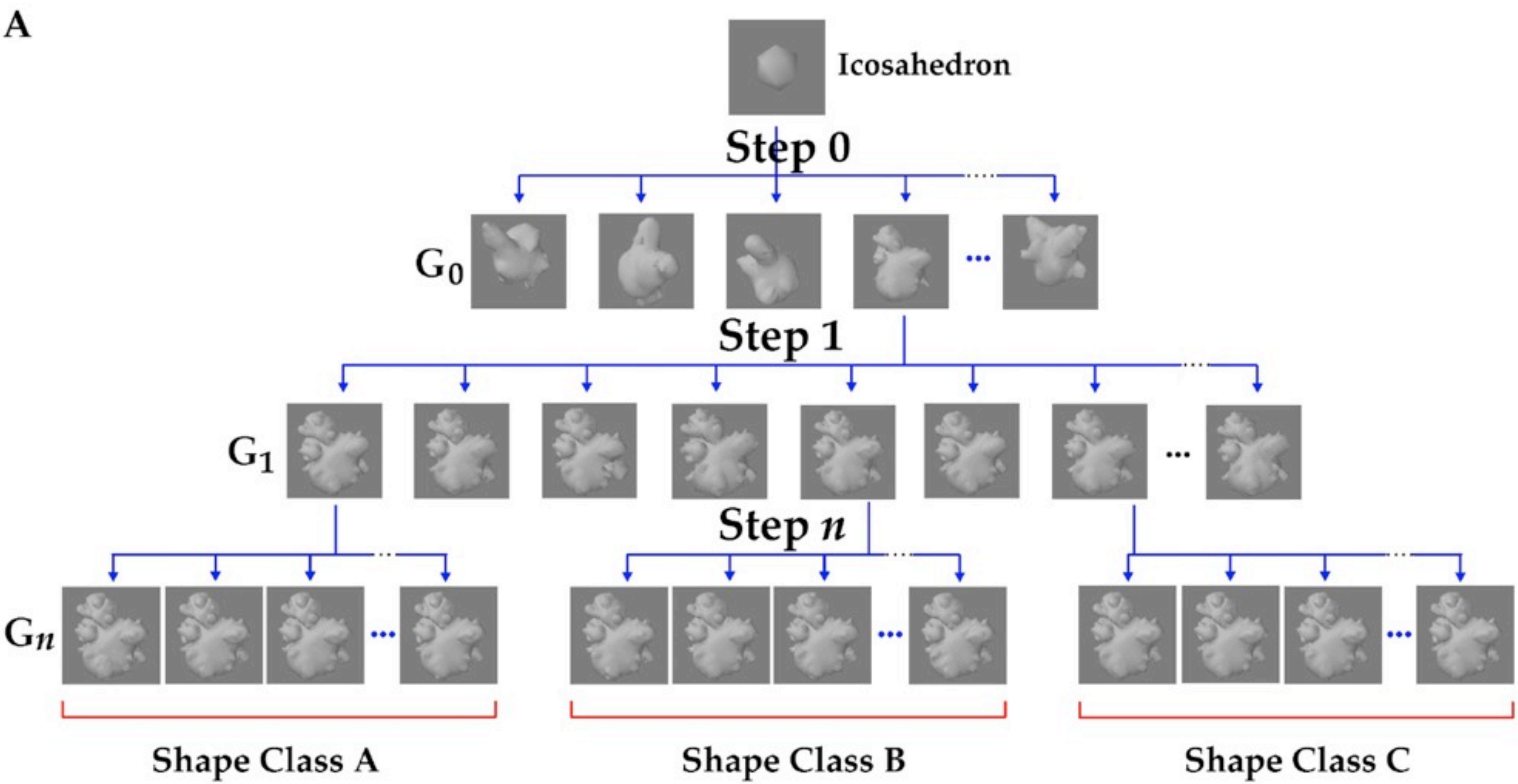
Digital embryo growth



Prof. Mark Brady

http://www.psych.ndsu.nodak.edu/brady/downloads.html

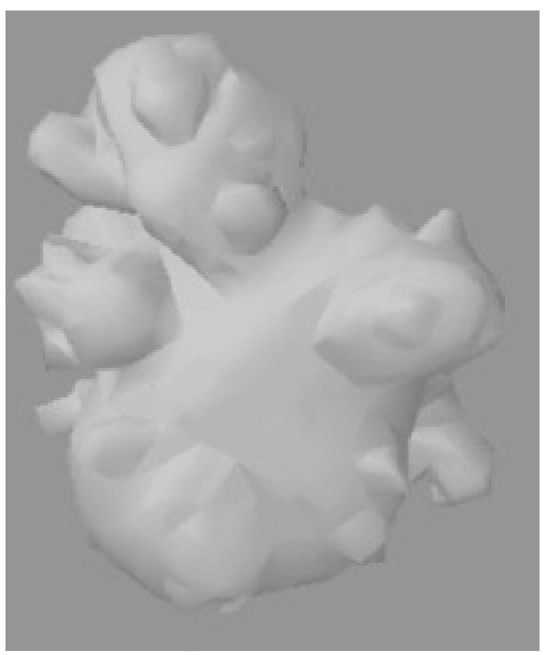
Virtual Phylogenesis

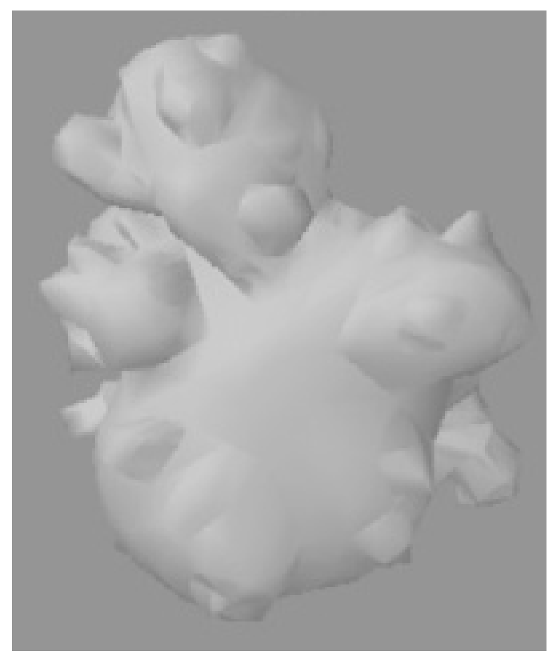


Training

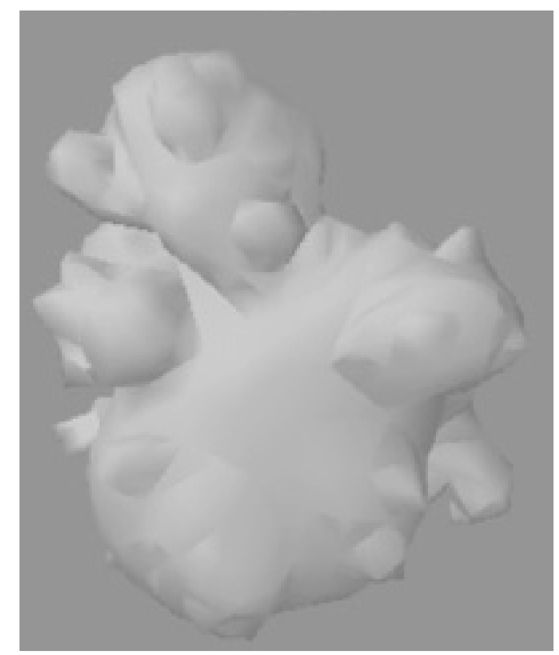
A or B?

A

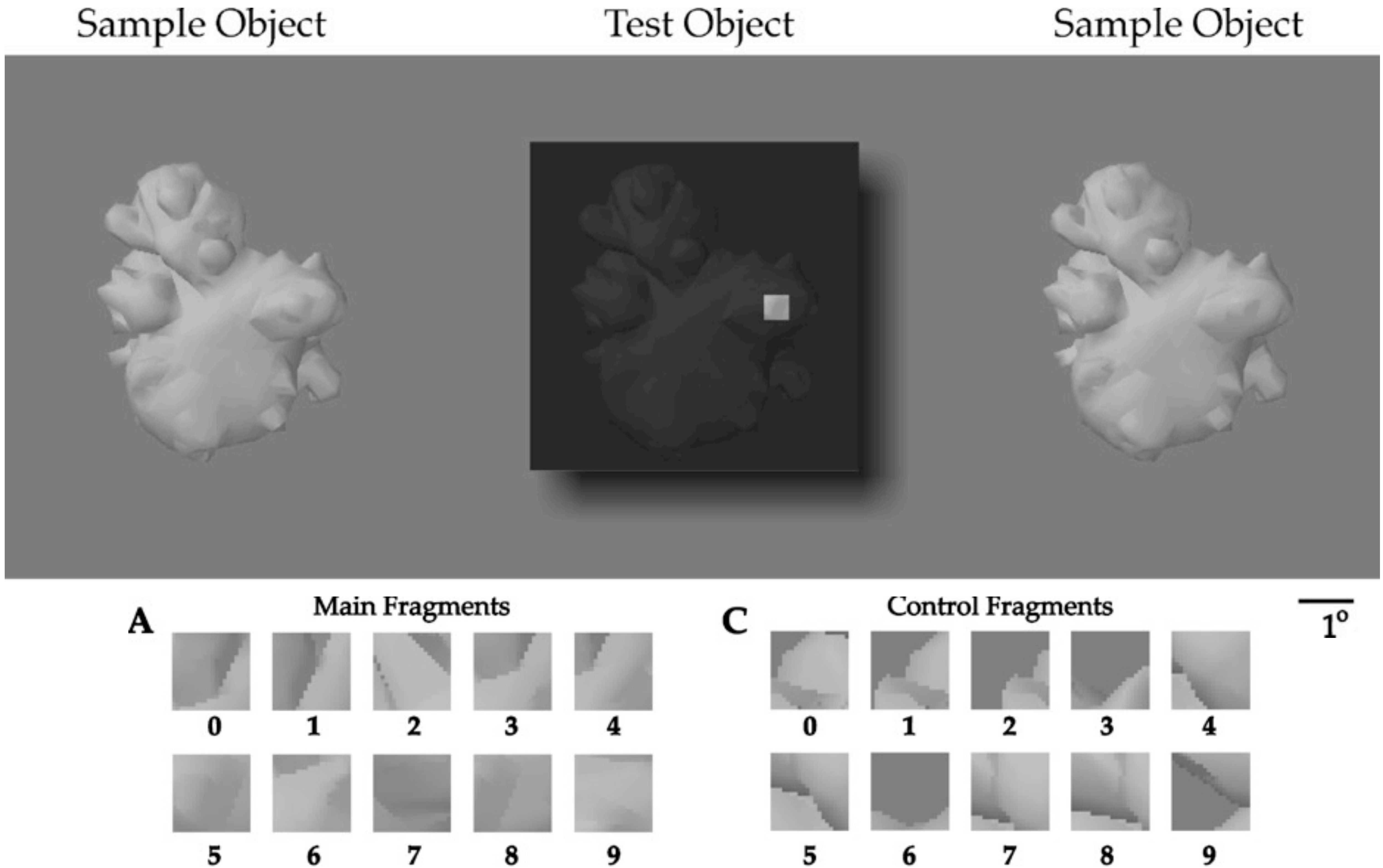




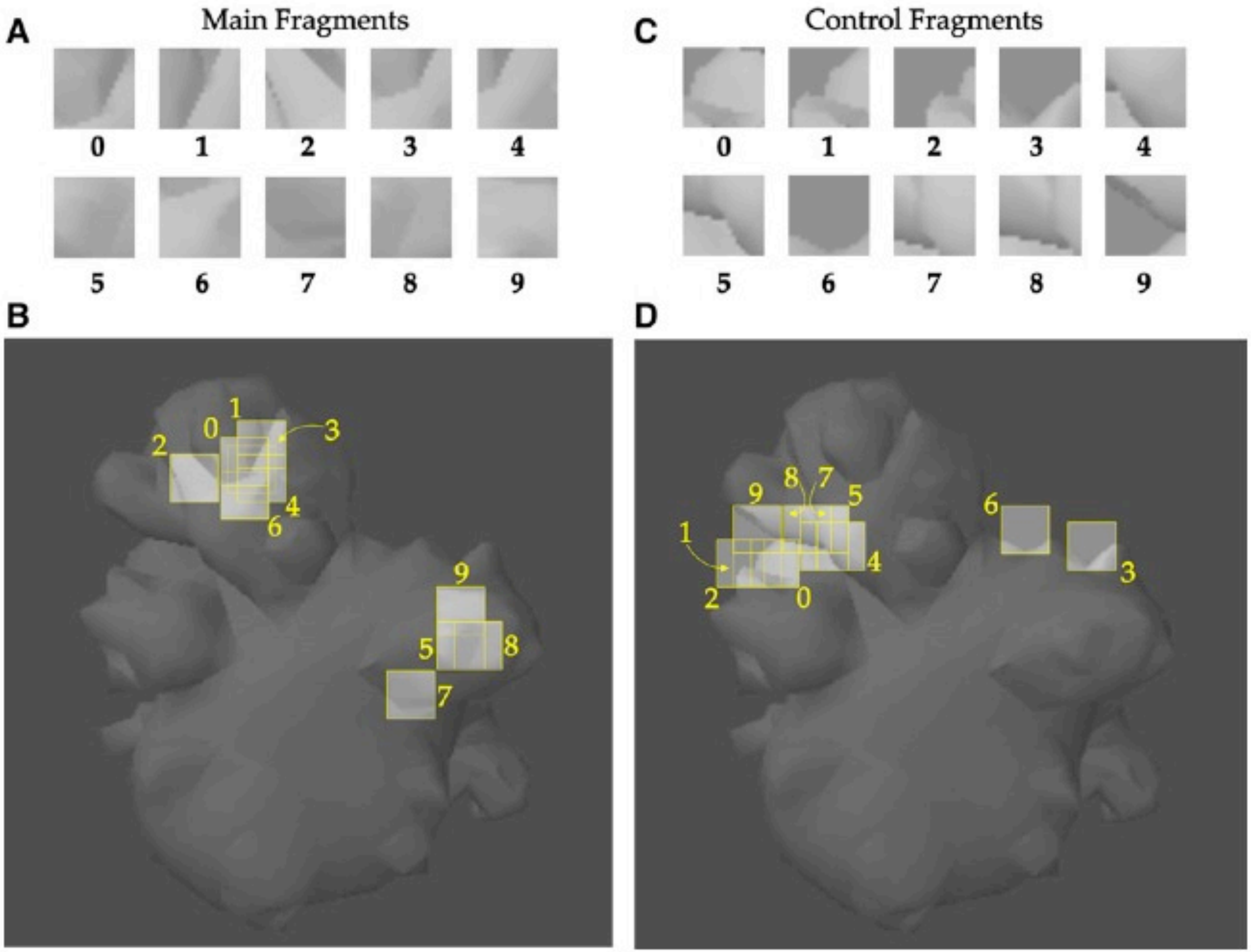
B



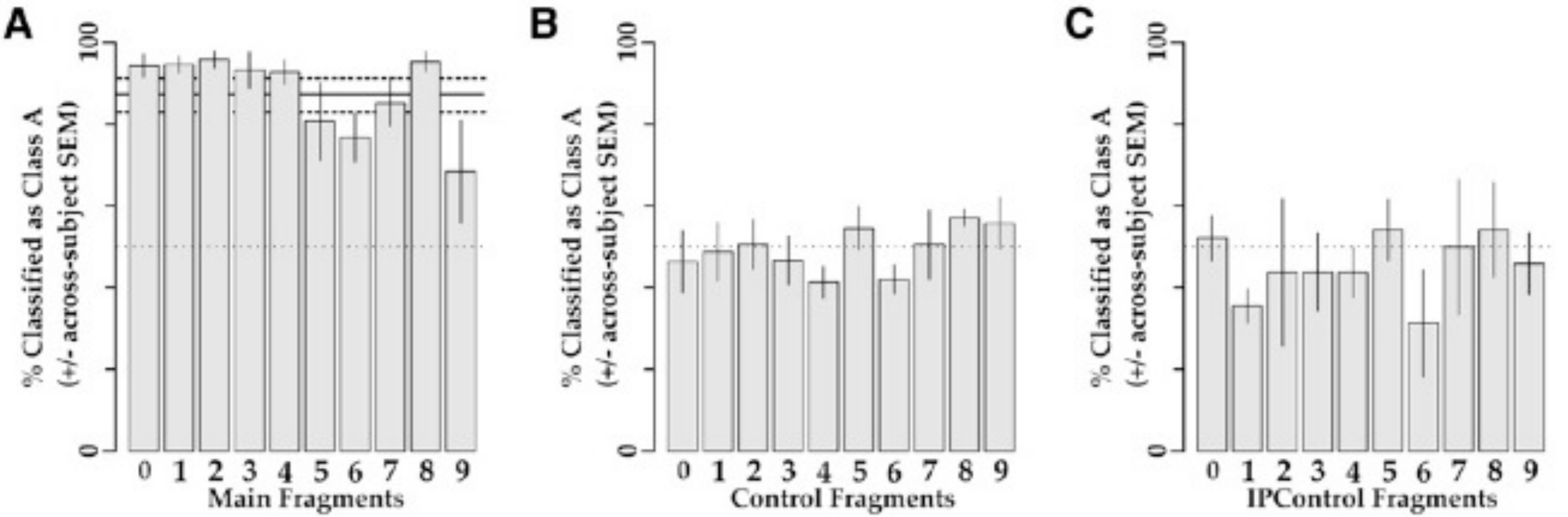
Testing


Sample Object
Test Object
Sample Object
A
Main Fragments
0
1
2
3
4
5
6
7
8
9
C
Control Fragments
0
1
2
3
4
5
6
7
8
9
1°

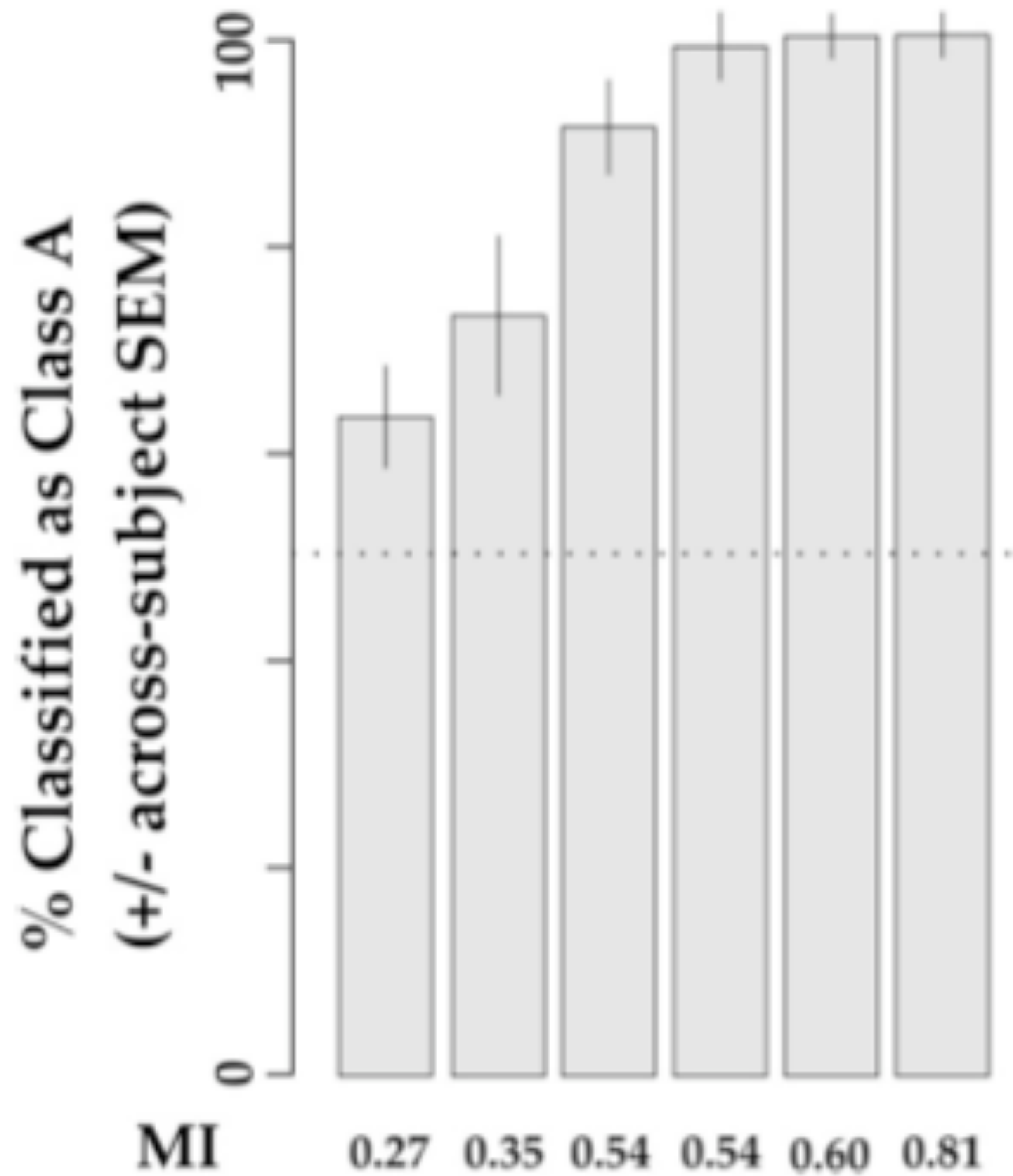
Fragments


A
Main Fragments
0
1
2
3
4
5
6
7
8
9
C
Control Fragments
0
1
2
3
4
5
6
7
8
9
B
D

Results

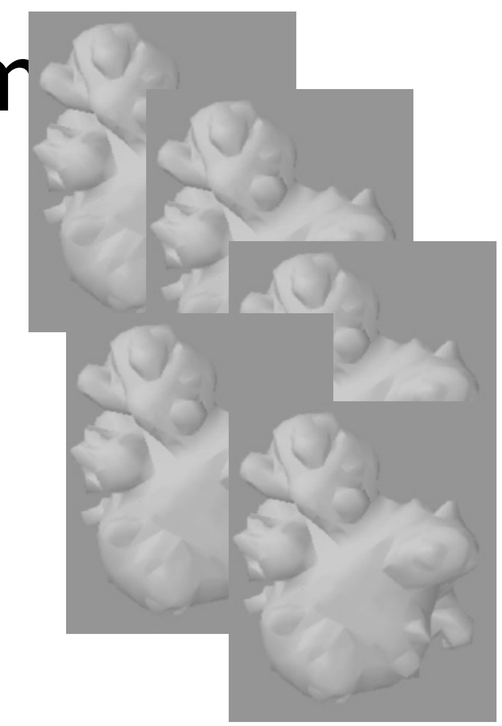


A
% Classified as Class A (+/- across-subject SEM)
100
0
0 1 2 3 4 5 6 7 8 9
Main Fragments
B
% Classified as Class A (+/- across-subject SEM)
100
0
0 1 2 3 4 5 6 7 8 9
Control Fragments
C
% Classified as Class A (+/- across-subject SEM)
100
0
0 1 2 3 4 5 6 7 8 9
IPControl Fragments


100
0
% Classified as Class A
(+/- across-subject SEM)
MI
0.27
0.35
0.54
0.54
0.60
0.81

Transfer of skill?

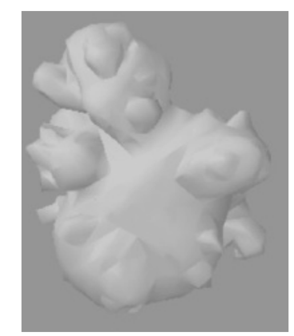
- For new previously unseem exemplars?
  - Yes. Maximizing mutual information seeks to provide an efficient set of features that are shared within a class, but at the sam time most effective at discriminating classes

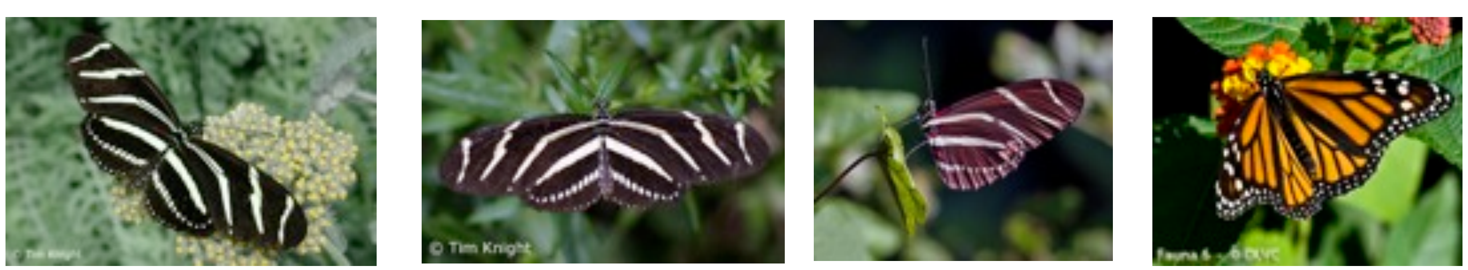


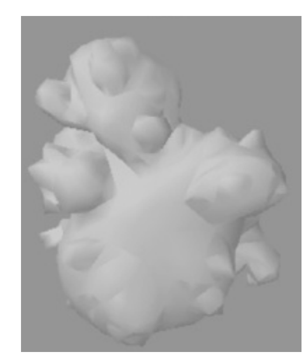
A

Transfer of skill?

- For new tasks that can be supported by the same discriminative features?
  - Yes.

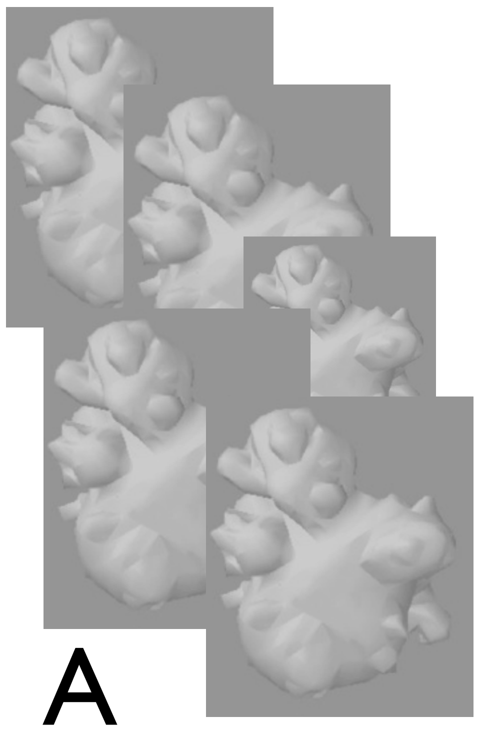
A  Big digitial embryo



B  Little digitial embryo

Classification training transfer to this?

Transfer of skill?

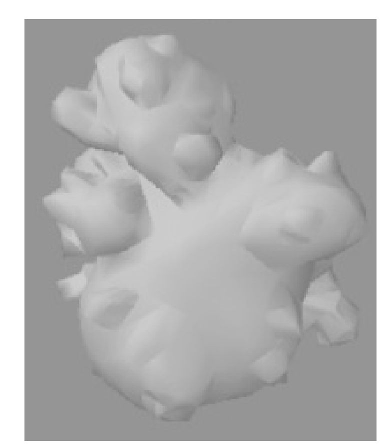


A

Learn A vs. B

B

Test A vs. C



C

Not effective transfer to new category C

Because humans learn to select the diagnostic features for A vs. B

General limitations

- Requires visual coherency
- Not straightforward to apply to conditions with clutter, background variations

Summing up

- Analysis-by-synthesis works best with good bottom-up processing
- Humans and machines need to learn diagnostic features that can rapidly and reliably support a variety of tasks
    - selecting features that maximize mutual information provide one way to do this