An Egocentric Perspective on Active Vision and Visual Object Learning in Toddlers

S. Bambach, D. Crandall, L. Smith, C. Yu.

ICDL 2017

Experiment presenters: Arjun, Ginevra

Their Experiments

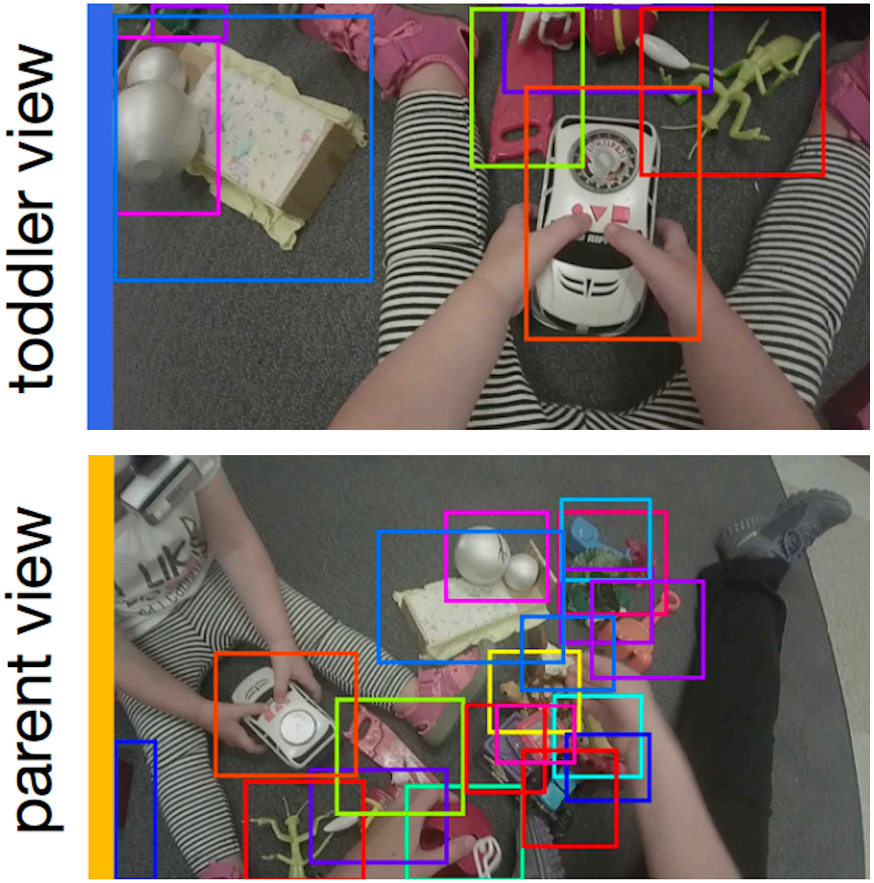
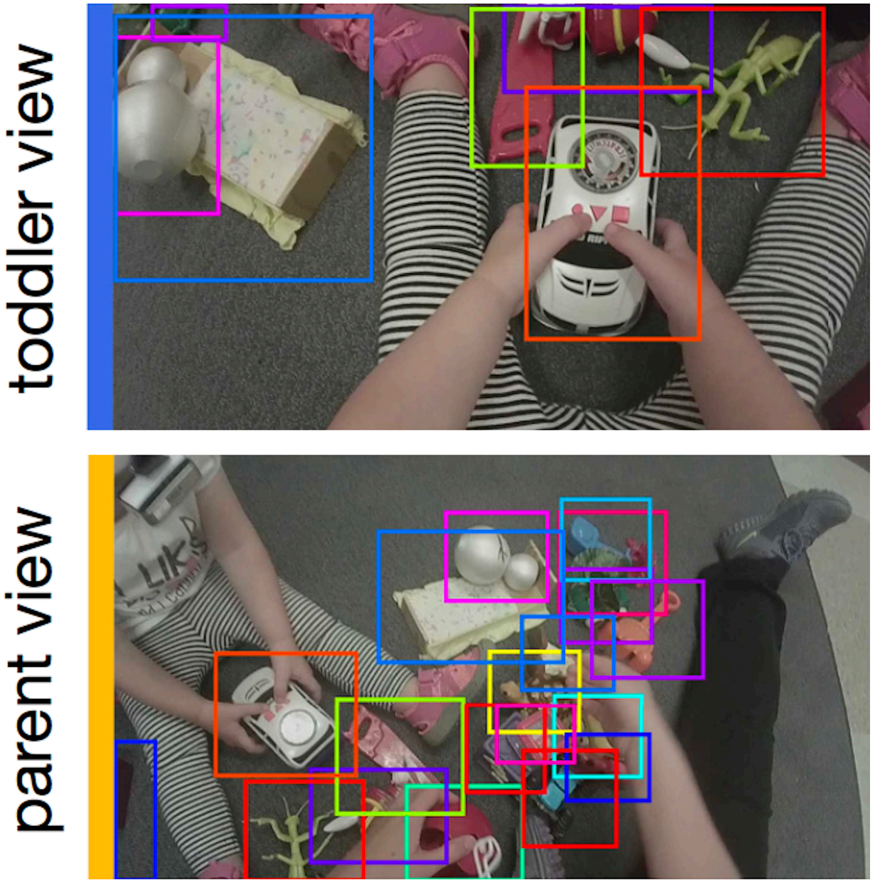


Image source: paper

Their Experiments

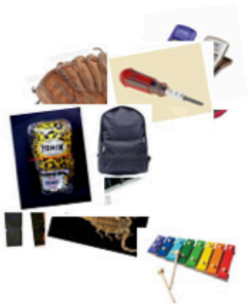


Authors could not control training set

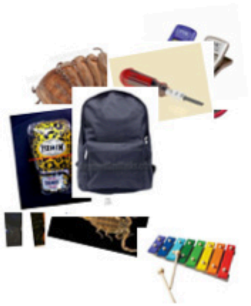
Image source: paper

Our Experiments

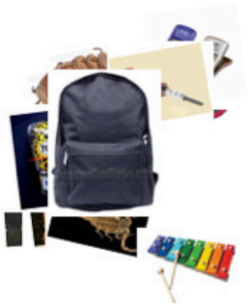
- We generate images where
  - – Labeled object occupies fixed percentage of view
  - – Background objects do not move



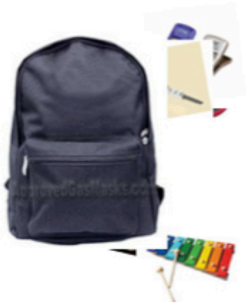
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



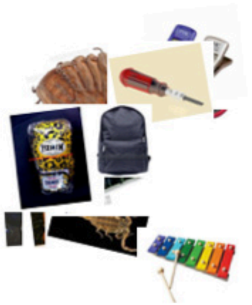
Scale 3
60% of view

Clean
Image

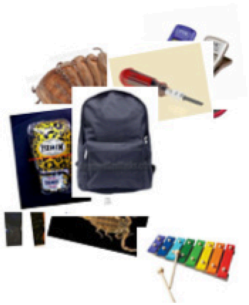
Image source: collages we made from Caltech 256 database

Our Experiments

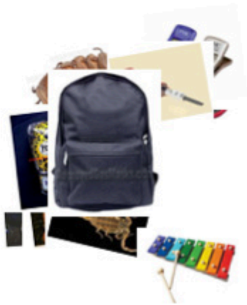
- Simulate toddler bringing object to face
  - We control scale to measure its effect on testing accuracy



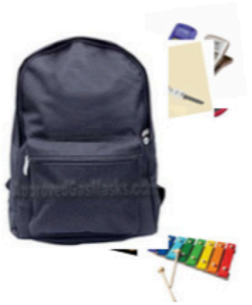
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



Scale 3
60% of view

Clean
Image

Image source: collages we made from Caltech 256 database

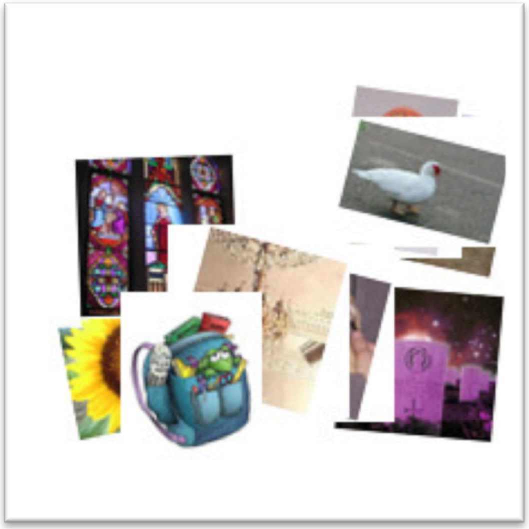
Our Dataset

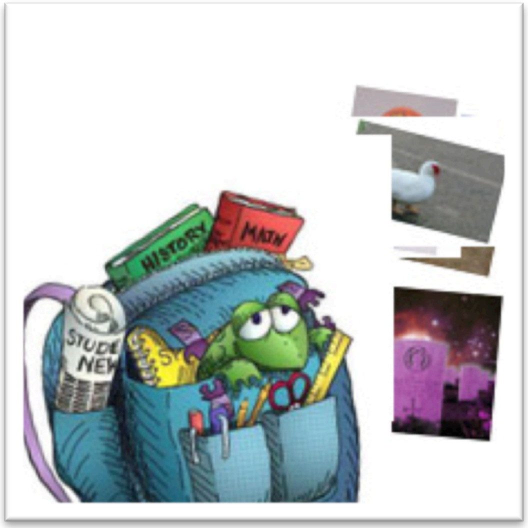
- 5 classes, 3633 images
- Collages
  - – Construct ‘scenes of toys’ using Caltech-256
  - – 1 positive image amongst many negatives
  - – Simulate toddler perspective

Image source: Caltech 256 database

Scene Generation

- Scene dim: **224 x 224**
  - Scale largest image dim to 70
  - Rotate randomly from -15° to 15°
- **10** negatives
  - Select uniformly from Caltech-256 negatives
  - Placed randomly in within scene boundary
- **1** positive
  - Scale 0 (1x), 1 (1.5x), 2 (2x), 3 (3x)
  - Place randomly within scene boundary (at scale 1)
- **2** scenes per training instance





VGG 16

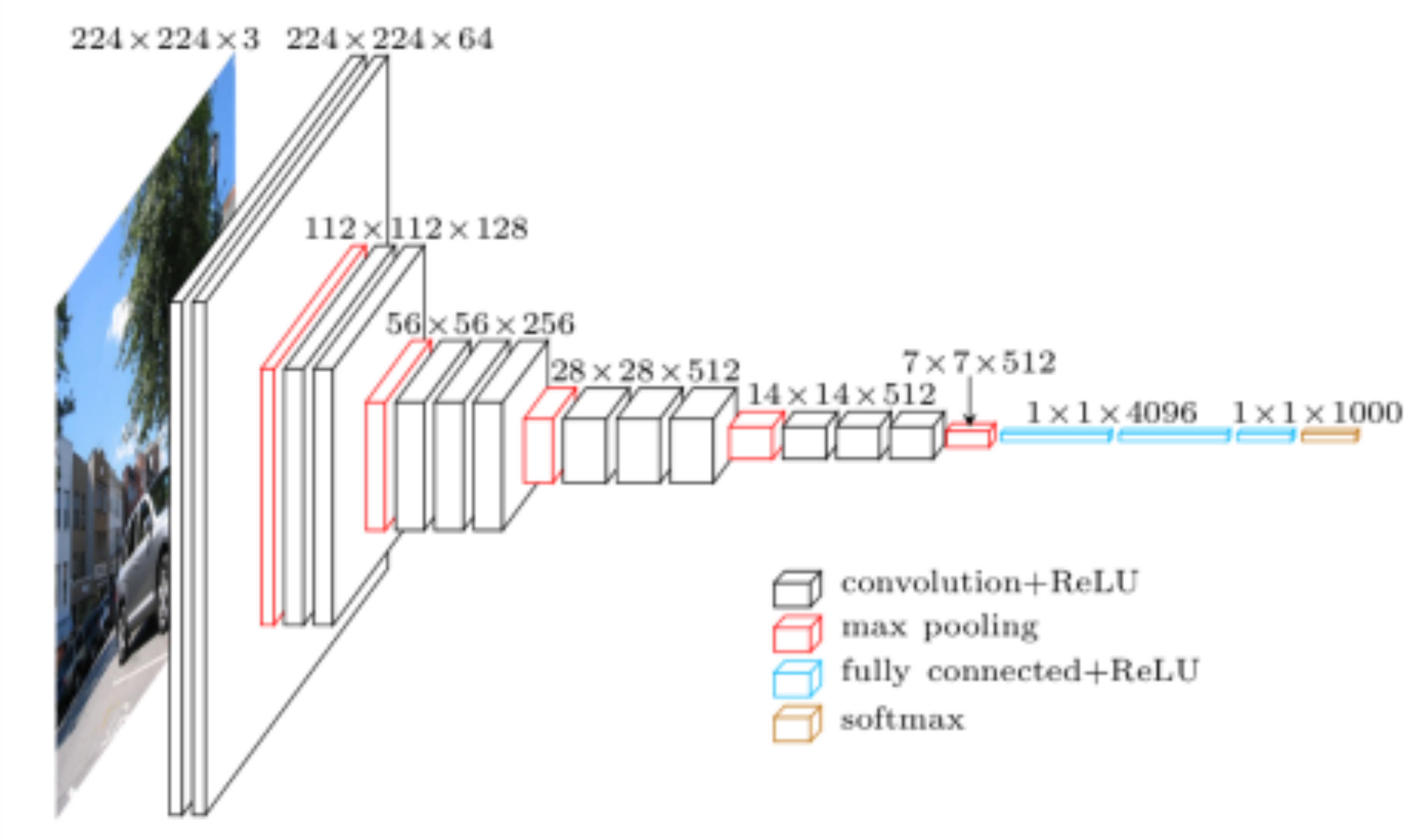


Image source, and source of some code used in the experiments: https://www.cs.toronto.edu/~frossard/post/vgg16/

VGG 16 for 5 classes

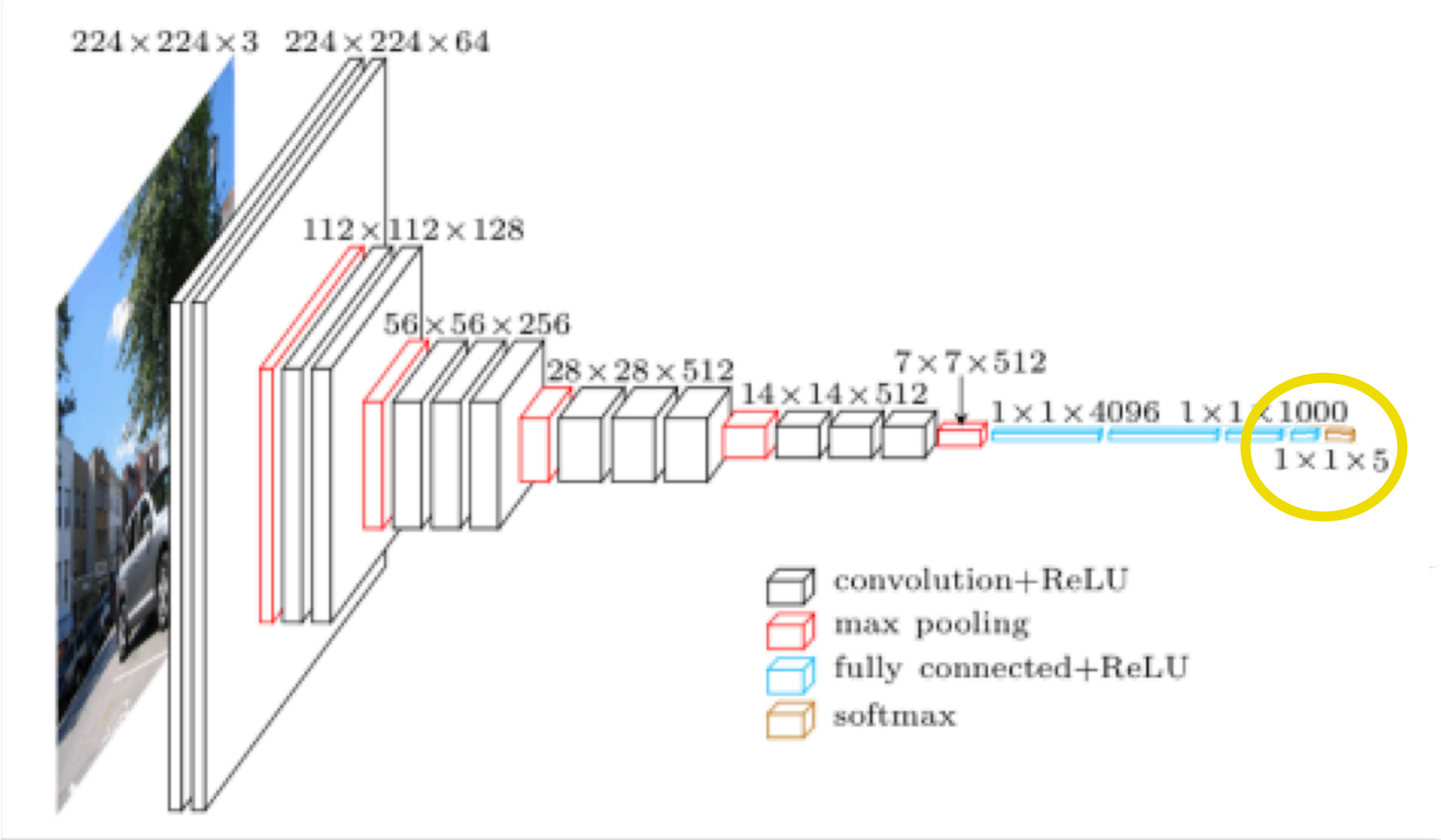
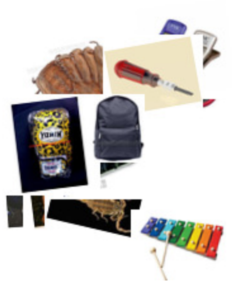


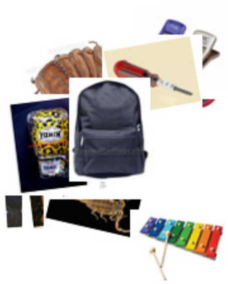
Image source: https://www.cs.toronto.edu/~frossard/post/vgg16/, modified by us

Experiment Setup

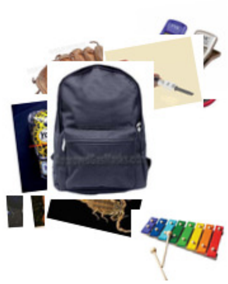
- Experiment 1
  - – Train on different scales, test on clean image
- Experiment 2
  - – Train on different scales and clean, test on different scales



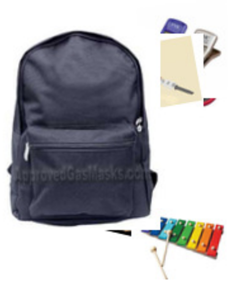
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



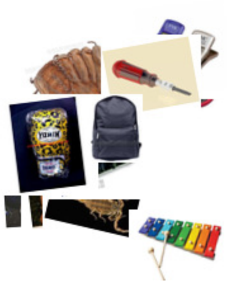
Scale 3
60% of view

Clean
Image

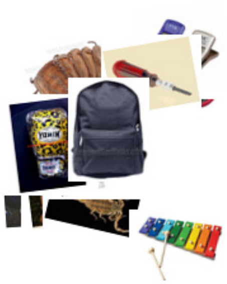
Image source: collages we made from Caltech 256 database

Experiment Setup

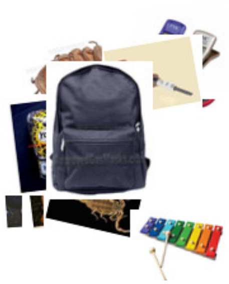
- Experiment 1
  - – Train on different scales, test on clean image
- Experiment 2
  - – Train on different scales and clean, test on different scales



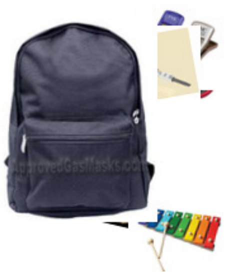
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



Scale 3
60% of view

Clean
Image

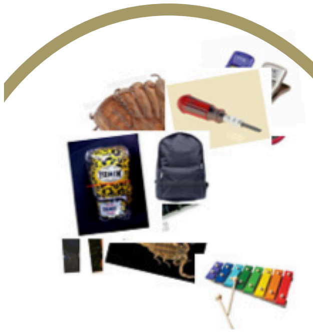
Image source: collages we made from Caltech 256 database

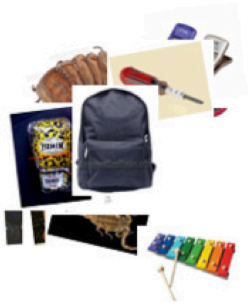
Experiment 1 - objective

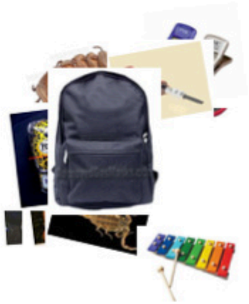
- Test effect of ‘bringing object to face’ for isolated classification
- Questions to consider
  - Effect of viewing at multiple scales?
  - Single ideal scale or result of multiple scales?

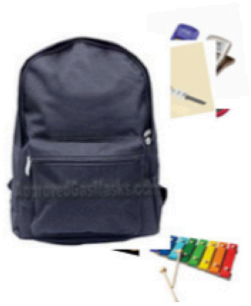
Image source: https://en.wiktionary.org/wiki/question_mark

Experiment 1 - data


Scale 0
10% of view


Scale 1
20% of view


Scale 2
30% of view

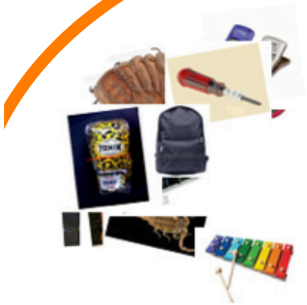

Scale 3
60% of view

Clean
Image

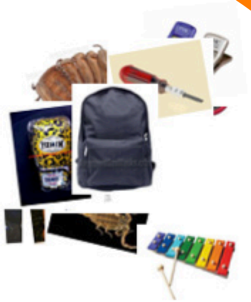
Train0

Image source: collages we made from Caltech 256 database

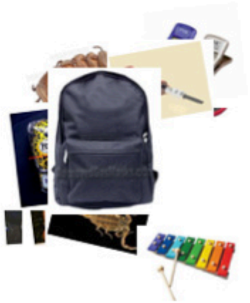
Experiment 1 - data



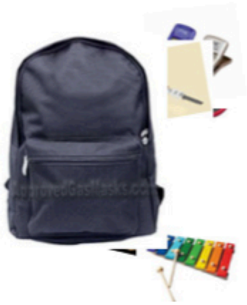
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



Scale 3
60% of view

Clean
Image

Train1

Image source: collages we made from Caltech 256 database

Experiment 1 - data

Scale 0
10% of view

Scale 1
20% of view

Scale 2
30% of view

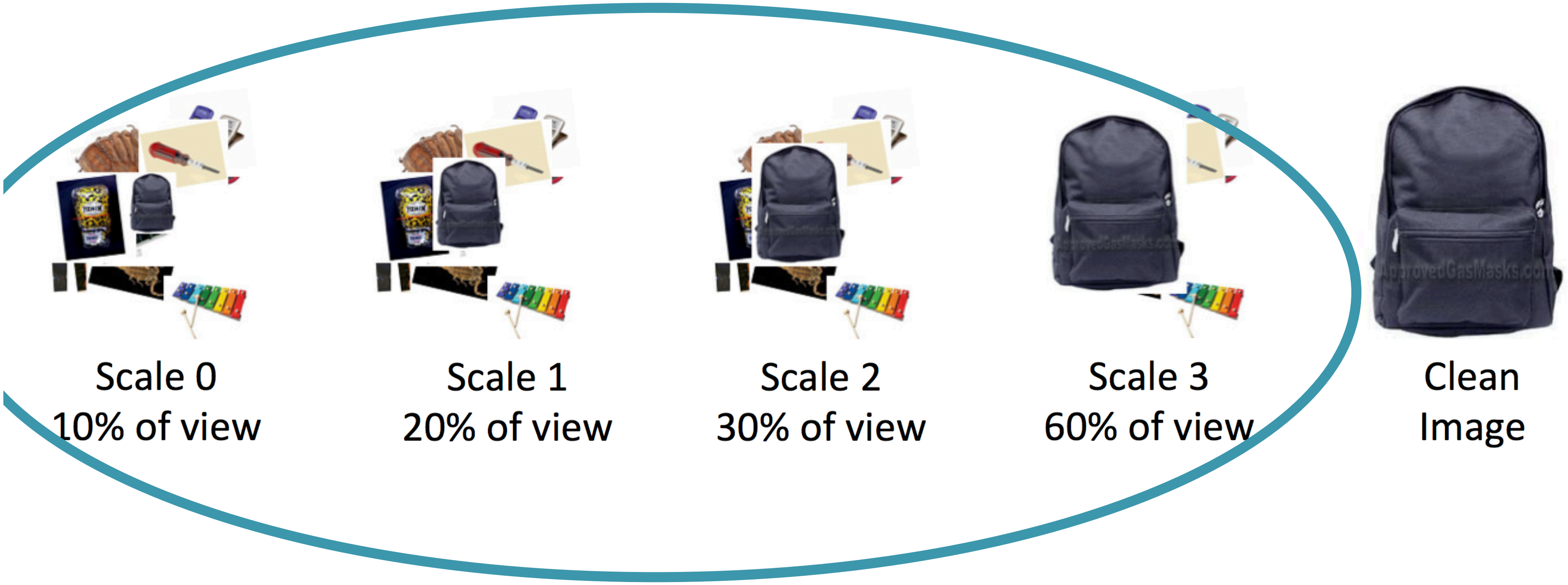
Scale 3
60% of view

Clean
Image

Train2

Image source: collages we made from Caltech 256 database

Experiment 1 - data



Scale 0
10% of view

Scale 1
20% of view

Scale 2
30% of view

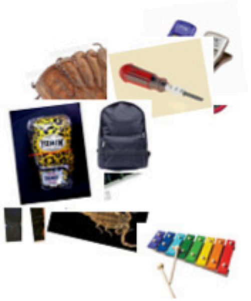
Scale 3
60% of view

Clean
Image

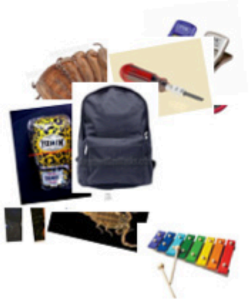
Train3

Image source: collages we made from Caltech 256 database

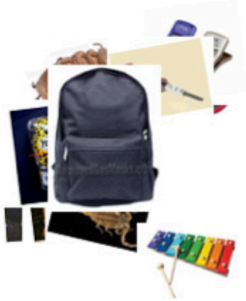
Experiment 1 - data



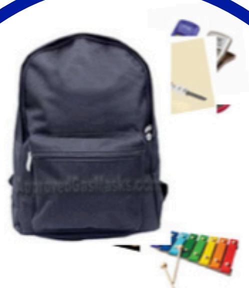
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



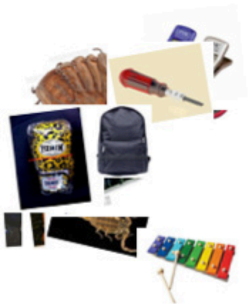
Scale 3
60% of view

Clean
Image

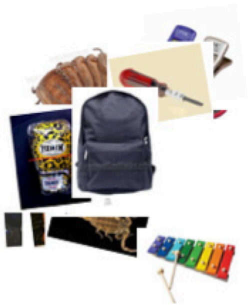
Train3only

Image source: collages we made from Caltech 256 database

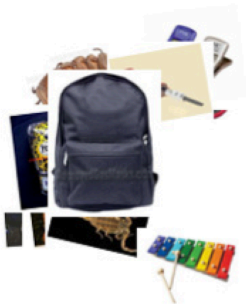
Experiment 1 - data



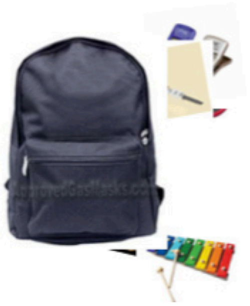
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



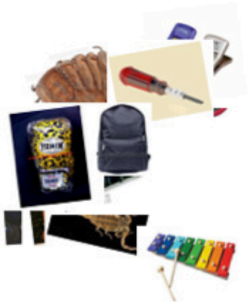
Scale 3
60% of view

Clean
Image

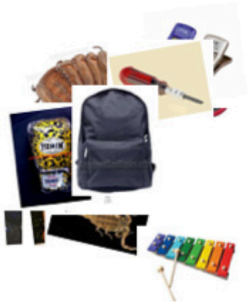
Correct number of epochs to compensate for more training examples

Image source: collages we made from Caltech 256 database

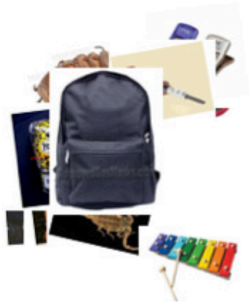
Experiment 1 - data



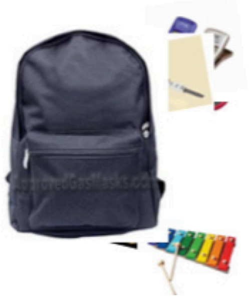
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



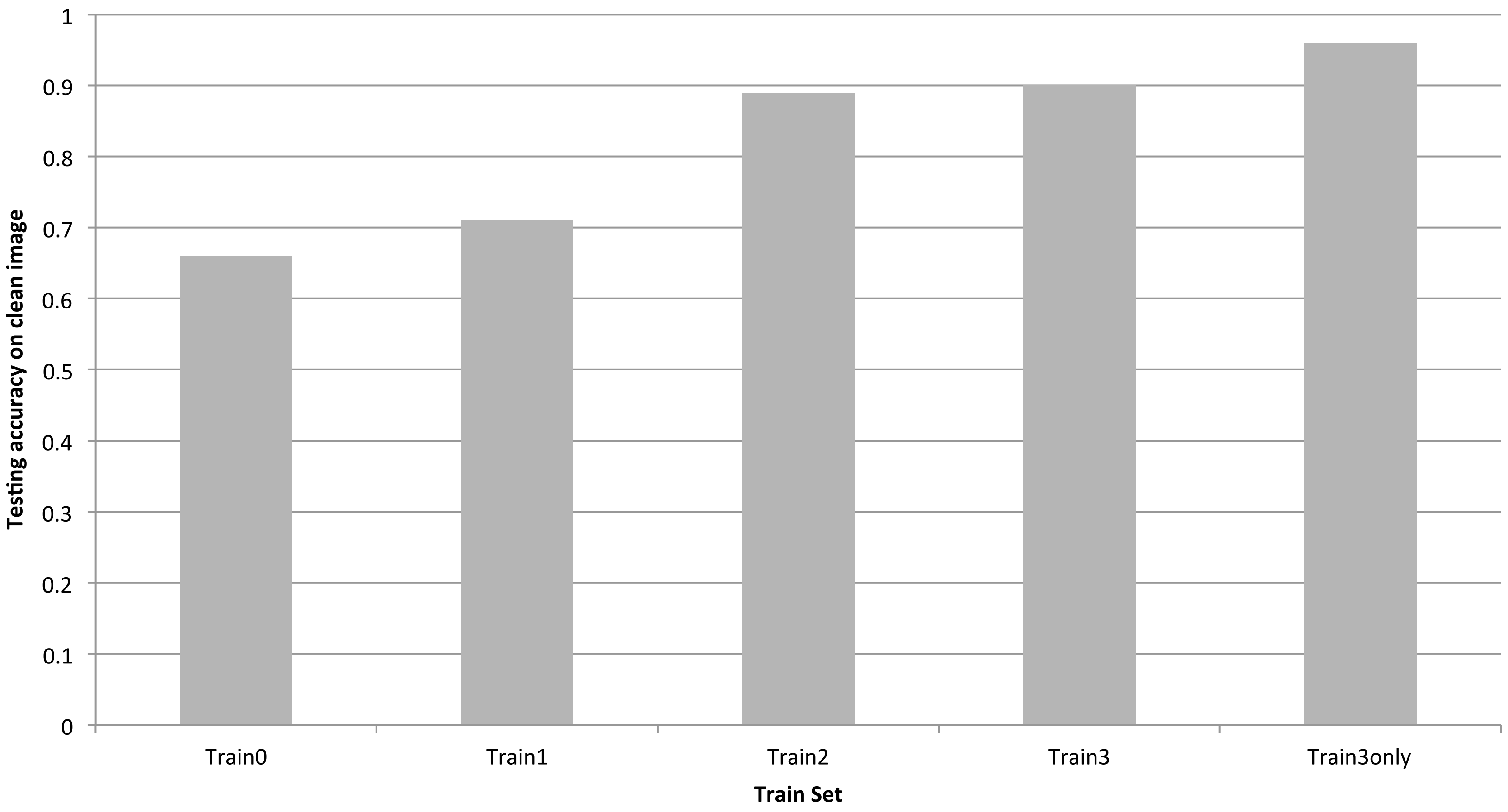
Scale 3
60% of view

Clean
Image

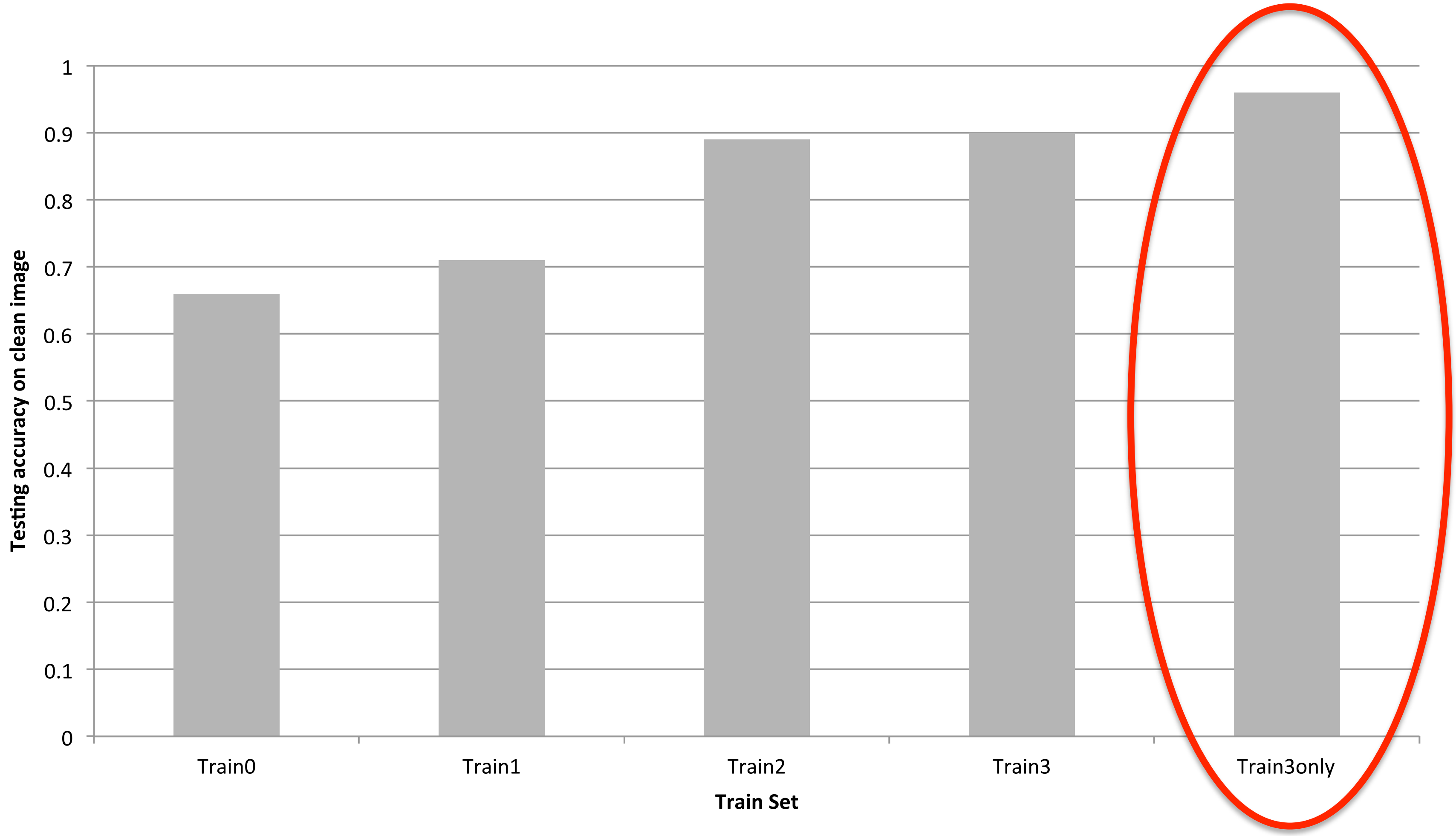
Test

Image source: collages we made from Caltech 256 database

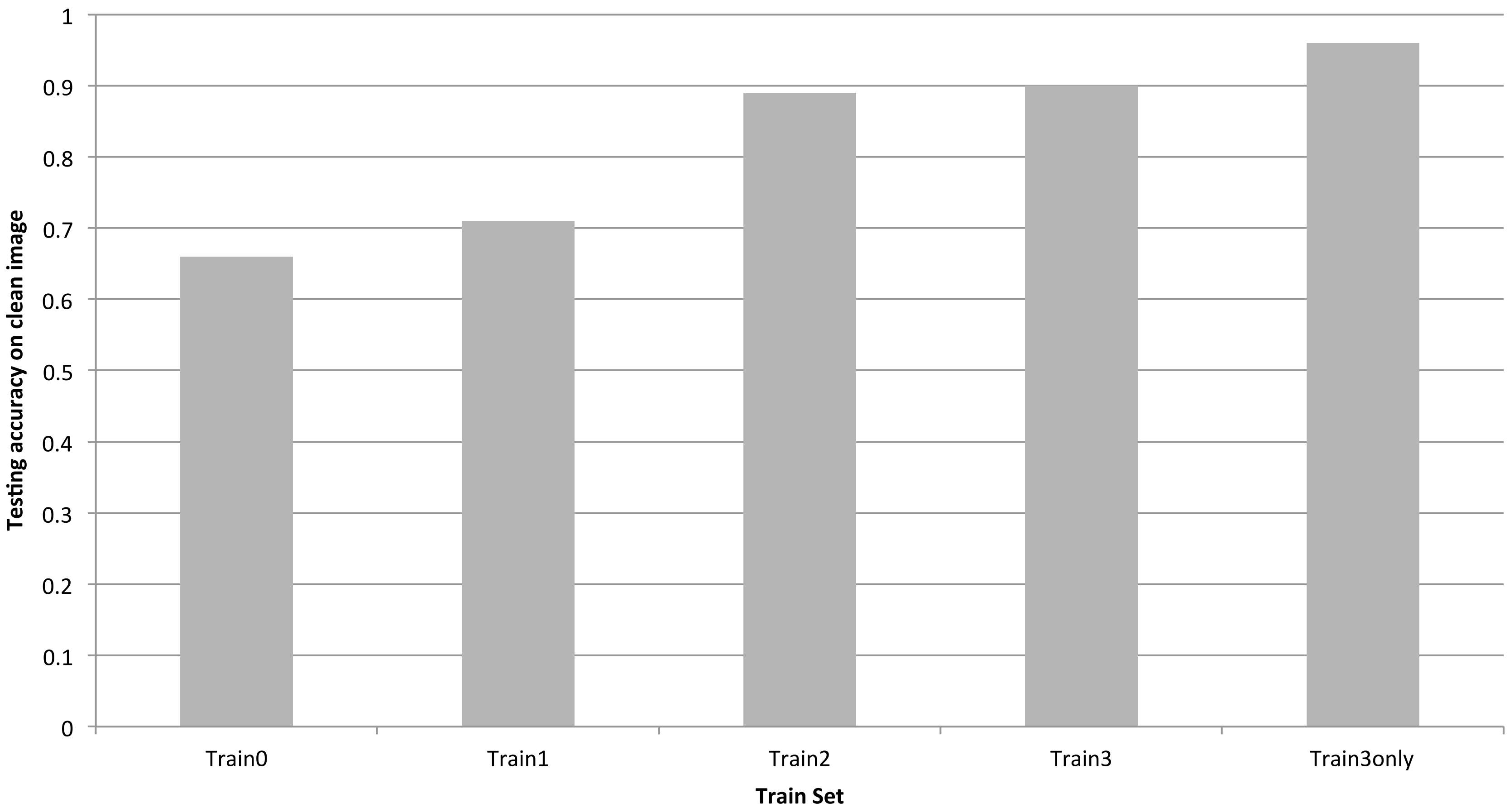
Experiment 1 - results



Experiment 1 - results



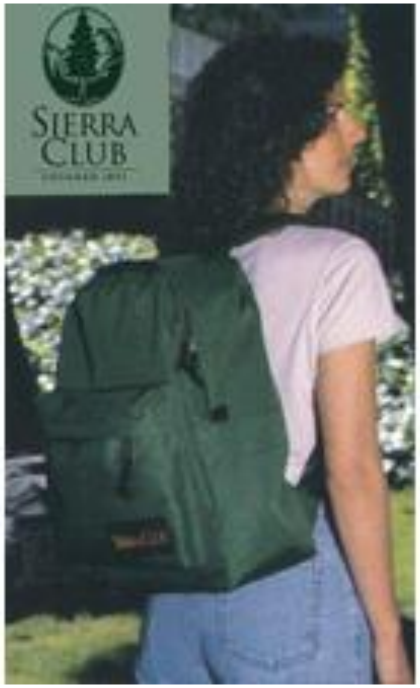
Experiment 1 - results



Training on larger scale images only yields to best test accuracy.

Experiment 1 - results

- Images misclassified when network trained in low scales benefit from training in higher scales



Misclassified after train0, train1, train2

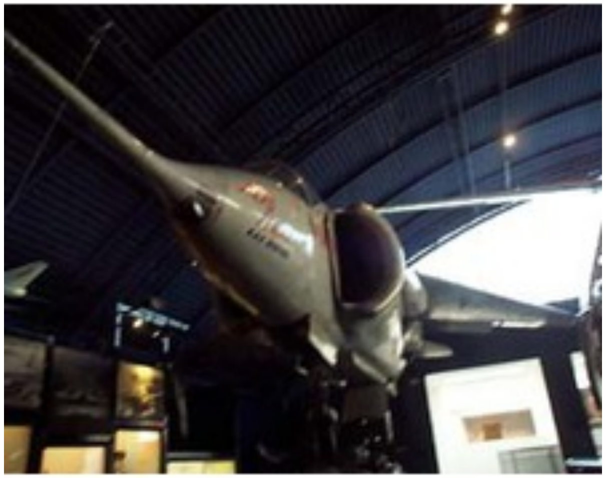
Correctly classified after train3 and train3only

(Category: bag)

Image source: Caltech 256 database

Experiment 1 - results

- Images misclassified when network trained in low scales benefit from training in higher scales



Misclassified after train0, train1, train2, train3

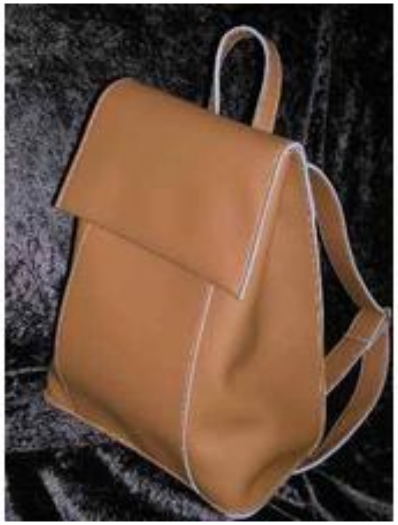
Correctly classified only after train3only

(Category: plane)

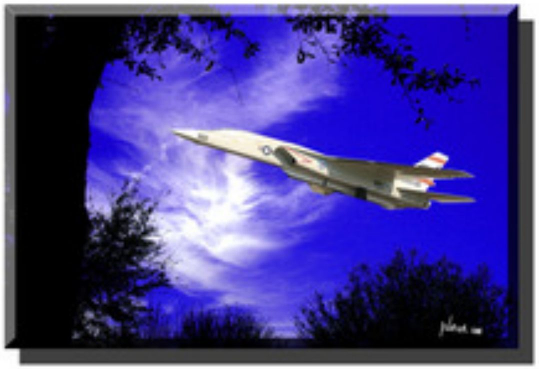
Image source: Caltech 256 database

Experiment 1 - results

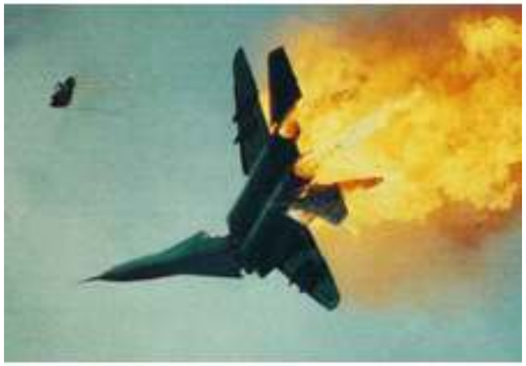
- Images misclassified after train3only were misclassified after all other trainings



Bag



Plane

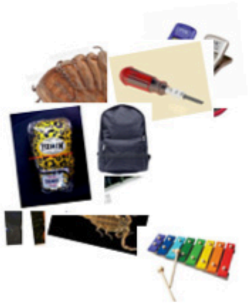


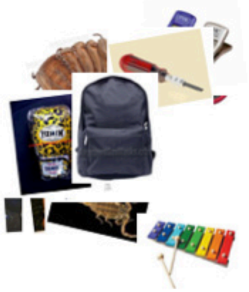
Plane

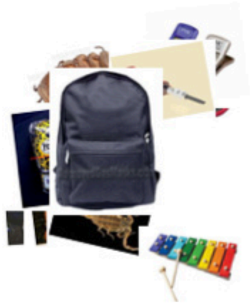
Image source: Caltech 256 database

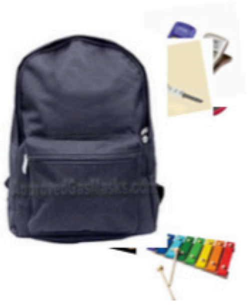
Experiment 1 - conclusions

- Toddler’s data gives better training because object is closer, not because it is ‘brought to face’
- Significant jump in accuracy if object occupies >30% of view in training
- Training images where object occupies <30% of view do more harm than good


Scale 0
10% of view


Scale 1
20% of view


Scale 2
30% of view

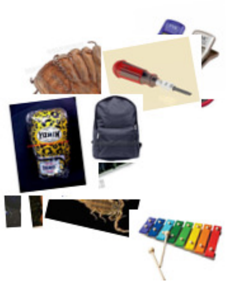

Scale 3
60% of view

Clean
Image

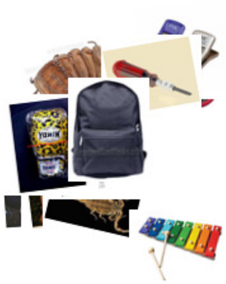
Image source: collages we made from Caltech 256 database

Experiment Setup

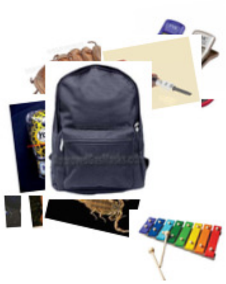
- Experiment 1
  - – Train on different scales, test on clean image
- Experiment 2
  - – Train on different scales and clean, test on different scales



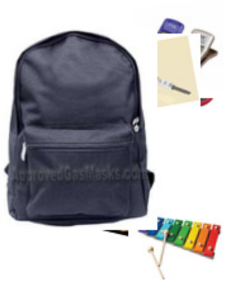
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



Scale 3
60% of view

Clean
Image

Image source: collages we made from Caltech 256 database

Experiment 2 - objective

- Effect of 'bringing to face' for object-in-scene detection
- Questions to consider
  - Does 'cleaning' the scene decrease detection in cluttered environment?

Image source: https://en.wiktionary.org/wiki/question_mark

Experiment 2 - data

Scale 0
10% of view

Scale 1
20% of view

Scale 2
30% of view

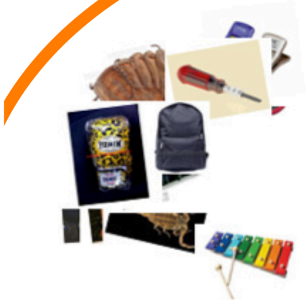
Scale 3
60% of view

Clean
Image

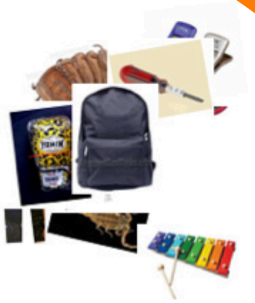
Train0

Image source: collages we made from Caltech 256 database

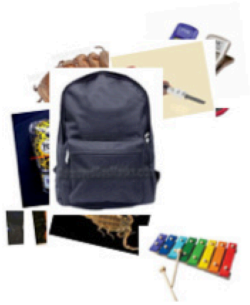
Experiment 2 - data



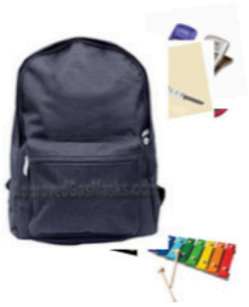
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



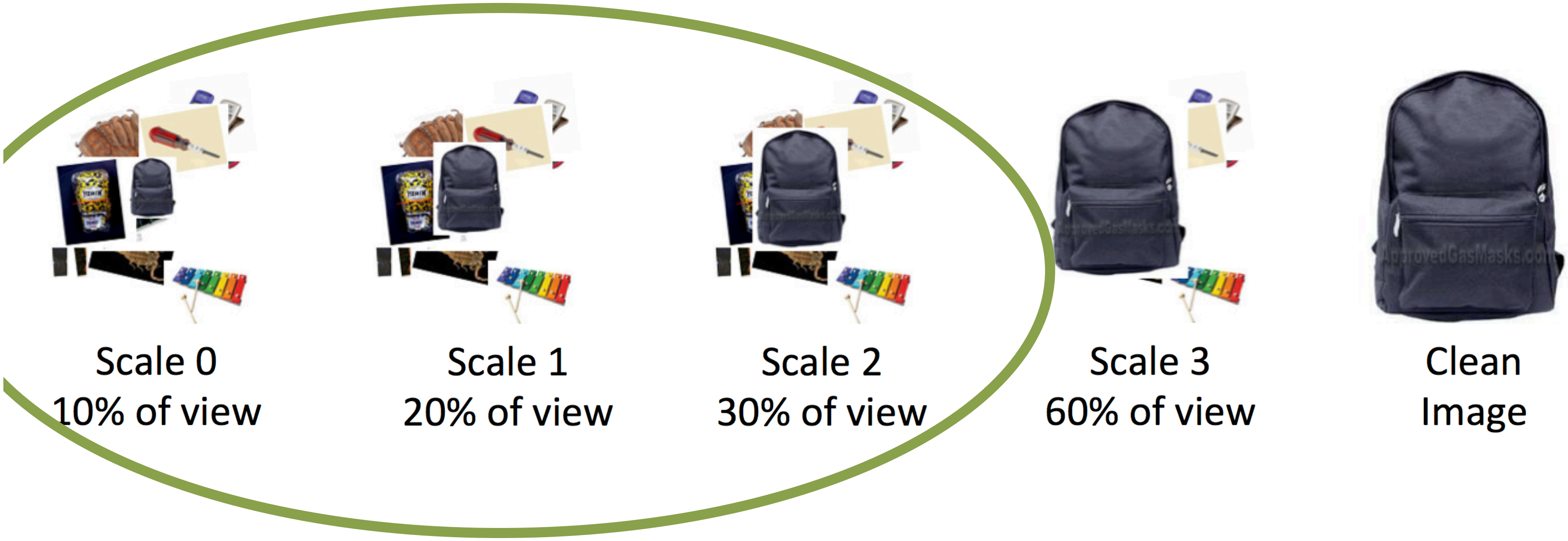
Scale 3
60% of view

Clean
Image

Train1

Image source: collages we made from Caltech 256 database

Experiment 2 - data



Scale 0
10% of view

Scale 1
20% of view

Scale 2
30% of view

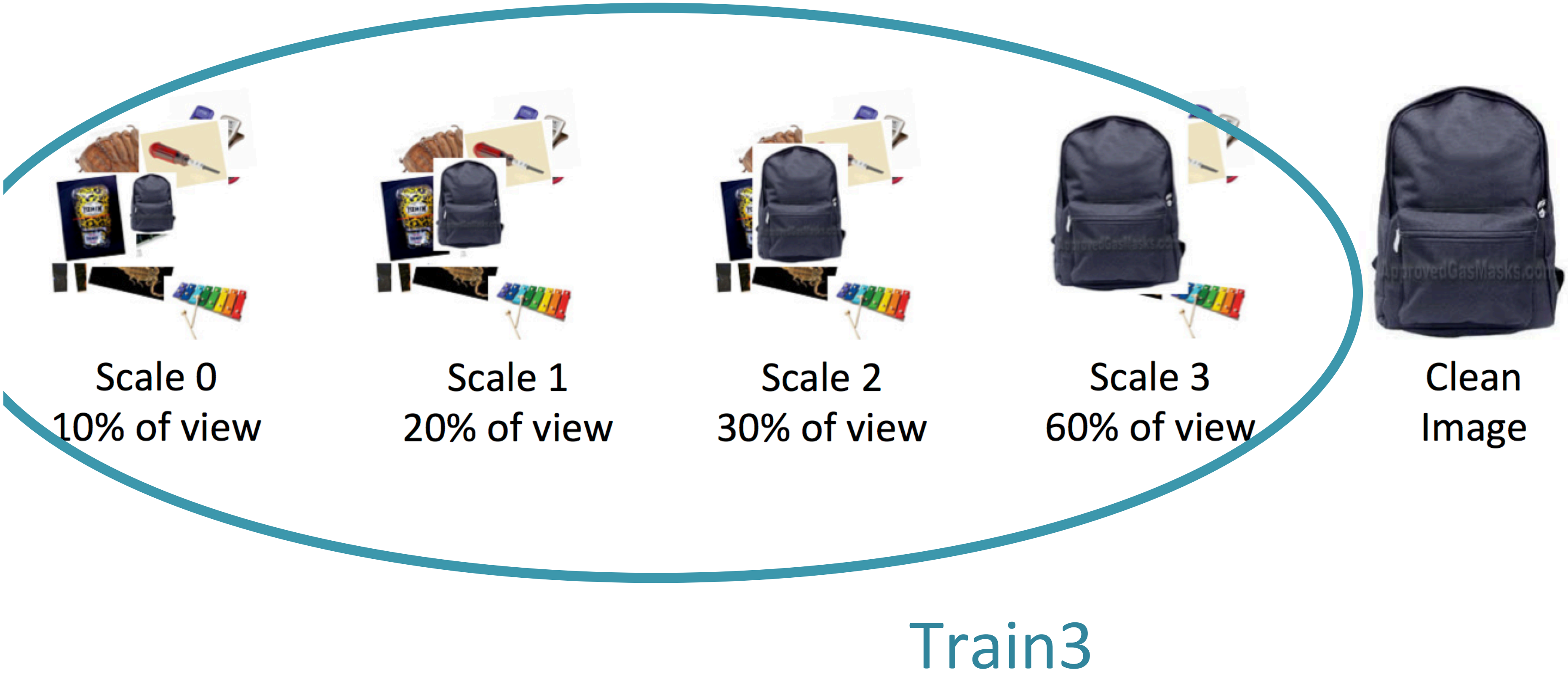
Scale 3
60% of view

Clean
Image

Train2

Image source: collages we made from Caltech 256 database

Experiment 2 - data



Scale 0
10% of view

Scale 1
20% of view

Scale 2
30% of view

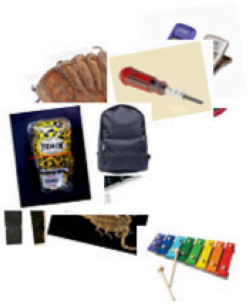
Scale 3
60% of view

Clean
Image

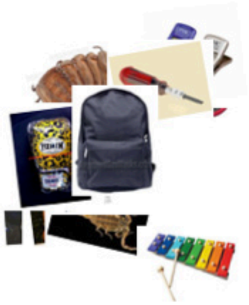
Train3

Image source: collages we made from Caltech 256 database

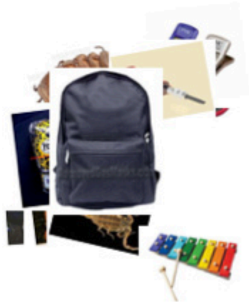
Experiment 2 - data



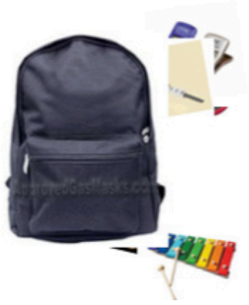
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



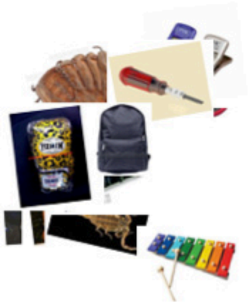
Scale 3
60% of view

Clean
Image

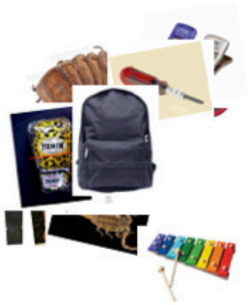
TrainClean

Image source: collages we made from Caltech 256 database

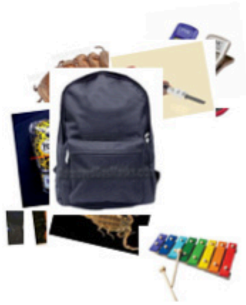
Experiment 2 - data



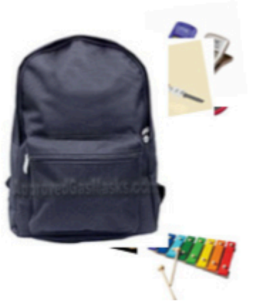
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



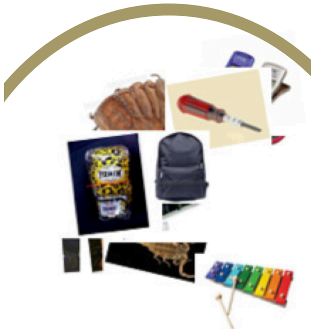
Scale 3
60% of view

Clean
Image

Correct number of epochs to compensate for more training examples

Image source: collages we made from Caltech 256 database

Experiment 2 - data



Scale 0
10% of view

Scale 1
20% of view

Scale 2
30% of view

Scale 3
60% of view

Clean
Image

Test0

On different images compared to train sets

Image source: collages we made from Caltech 256 database

Experiment 2 - data

Scale 0
10% of view

Scale 1
20% of view

Scale 2
30% of view

Scale 3
60% of view

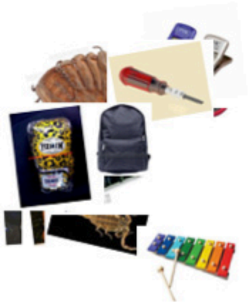
Clean
Image

Test1only

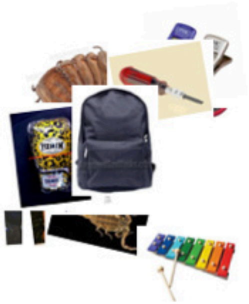
On different images compared to train sets

Image source: collages we made from Caltech 256 database

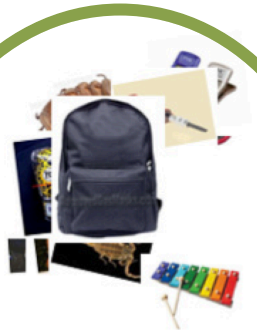
Experiment 2 - data



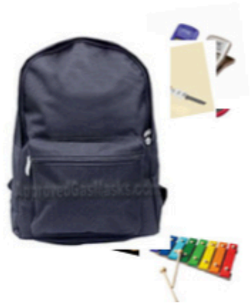
Scale 0
10% of view



Scale 1
20% of view



Scale 2
30% of view



Scale 3
60% of view

Clean
Image

Test2only

On different images compared to train sets

Image source: collages we made from Caltech 256 database

Experiment 2 - data

Scale 0
10% of view

Scale 1
20% of view

Scale 2
30% of view

Scale 3
60% of view

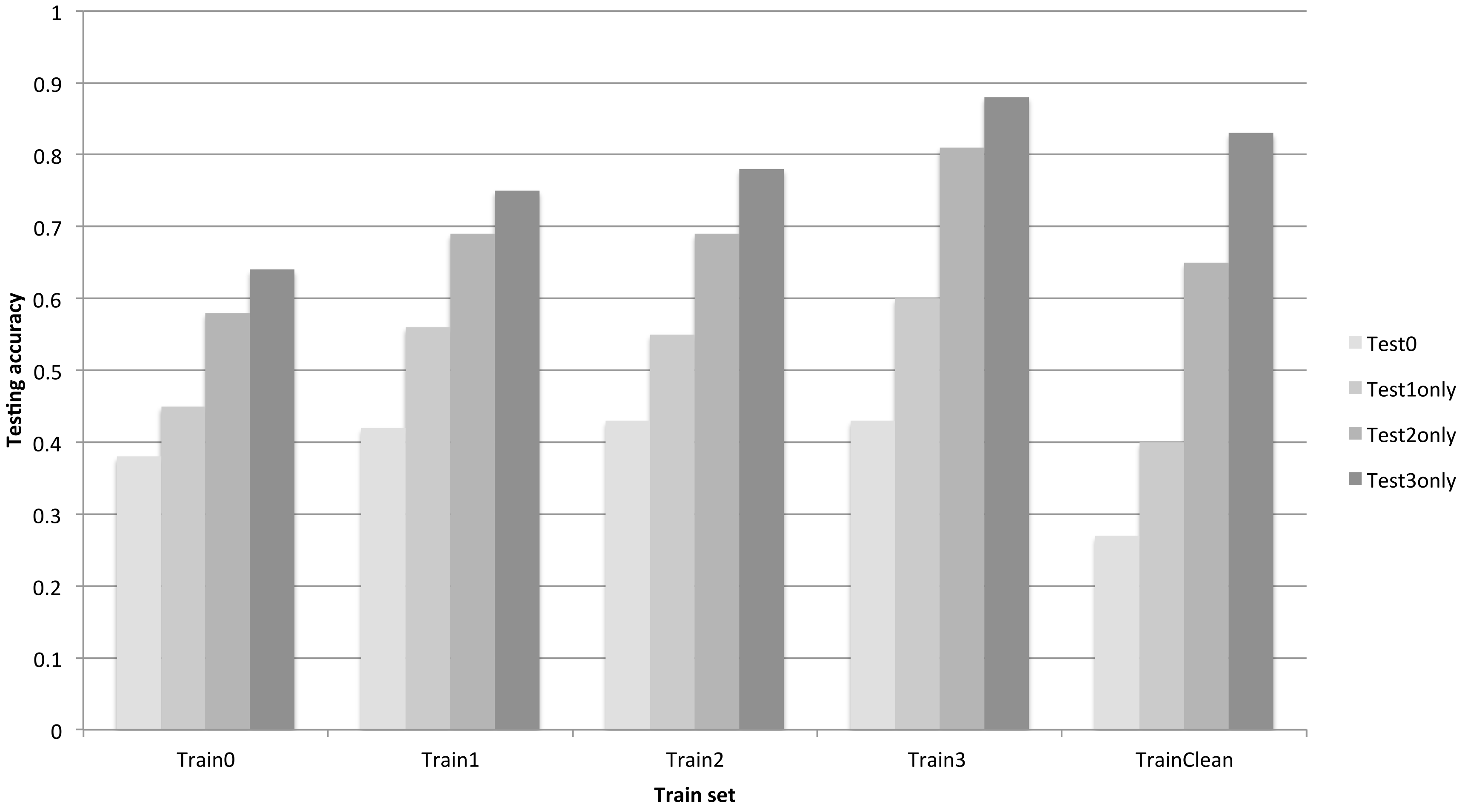
Clean
Image

Test3only

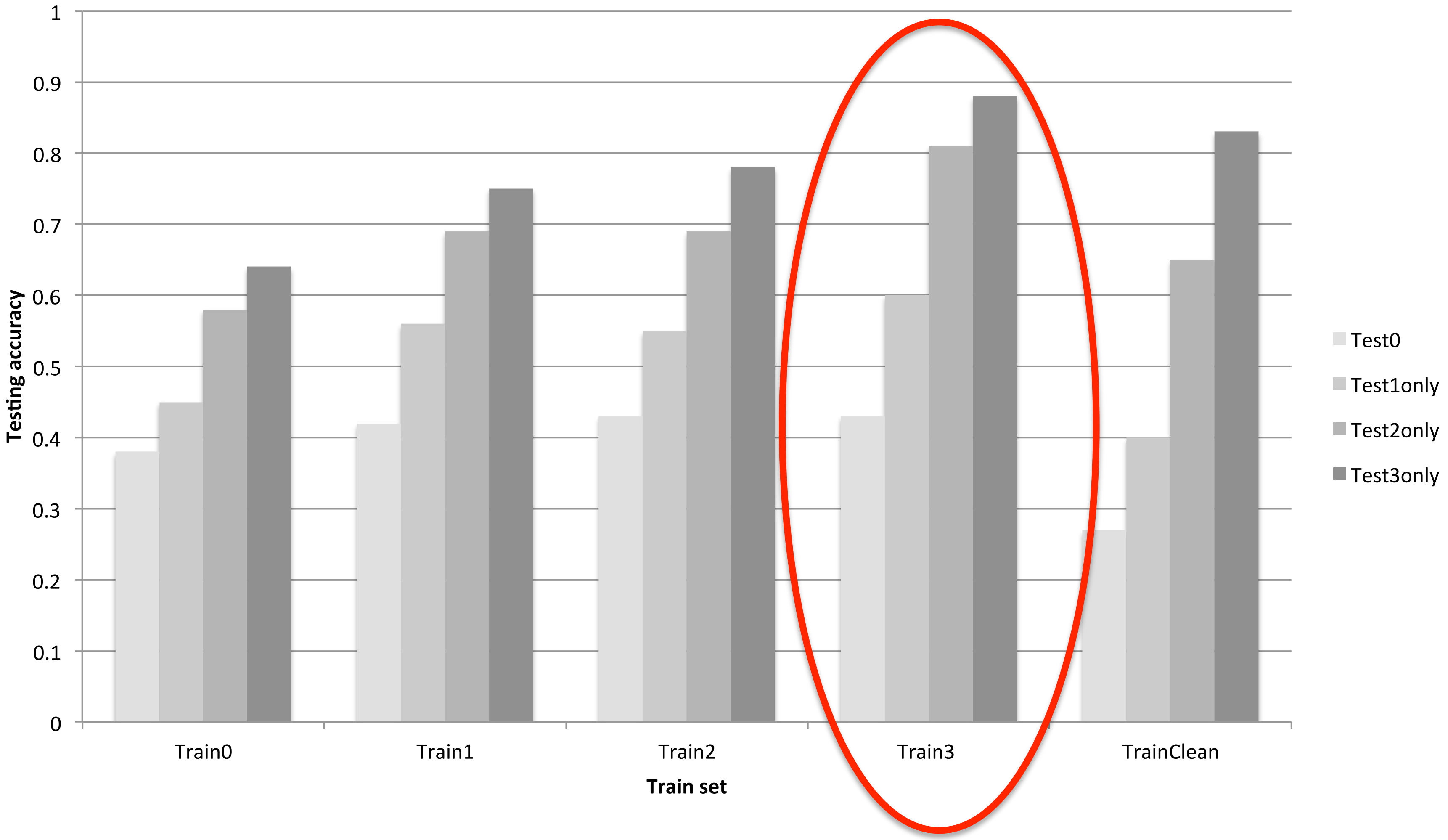
On different images compared to train sets

Image source: collages we made from Caltech 256 database

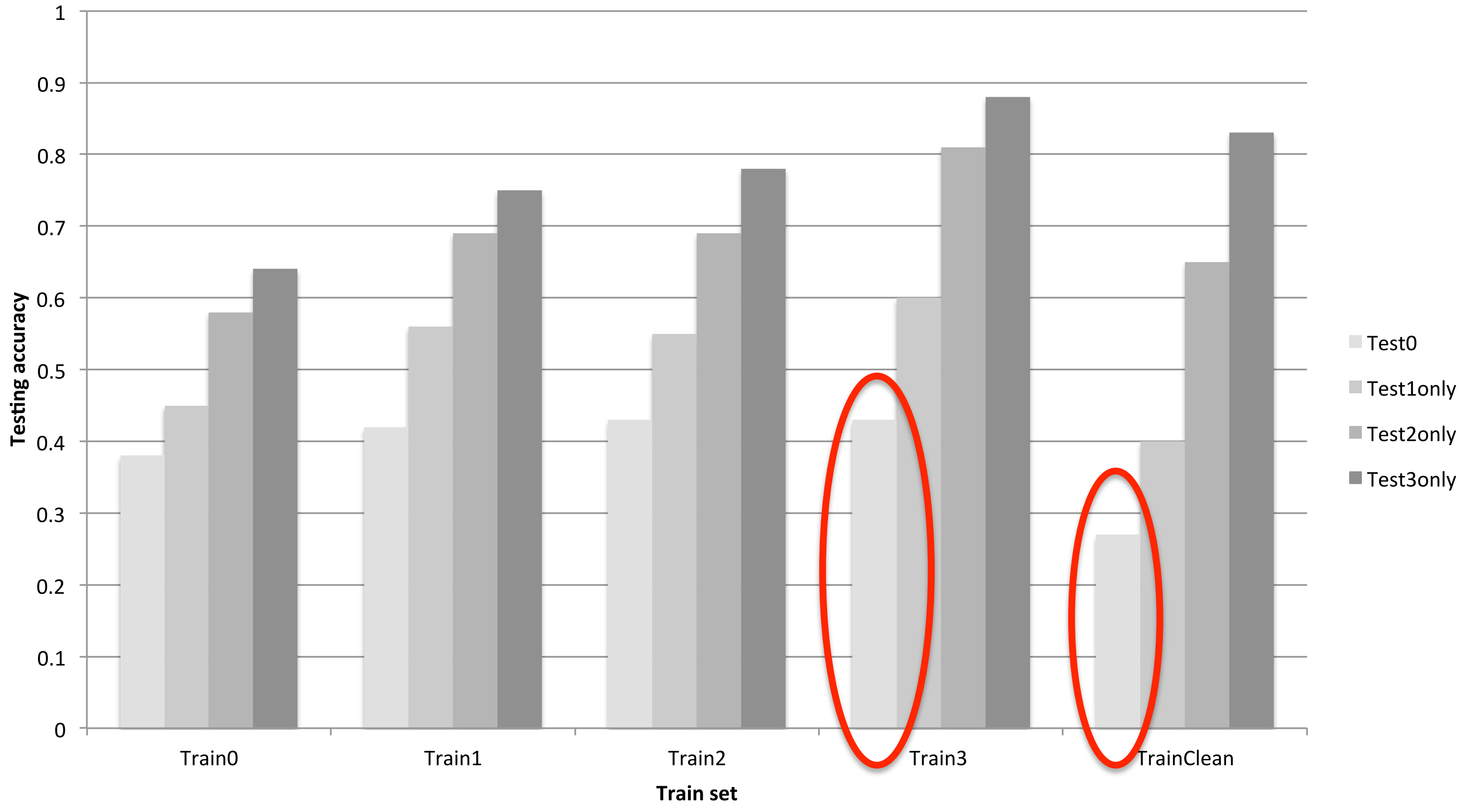
Experiment 2 - results



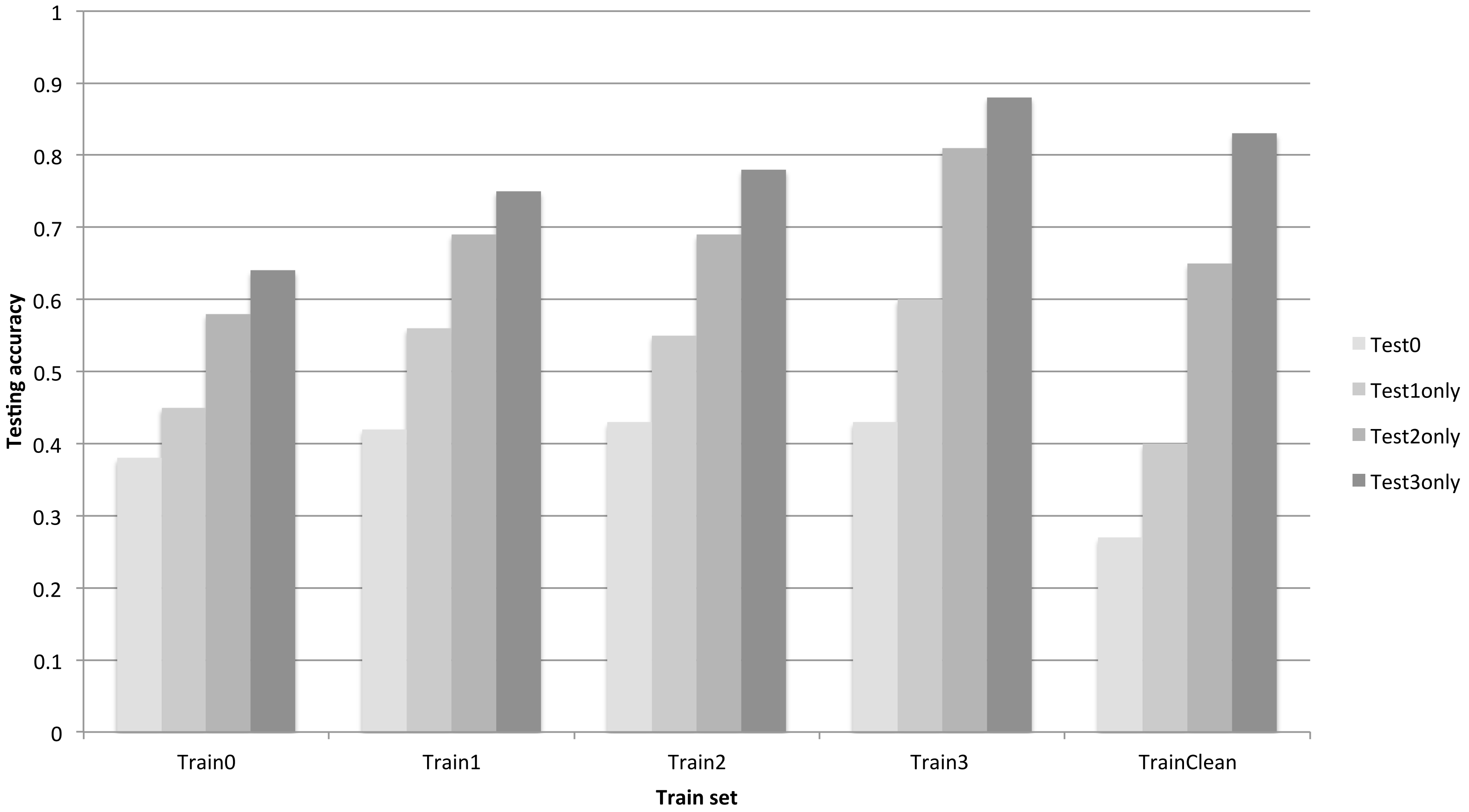
Experiment 2 - results



Experiment 2 - results



Experiment 2 - results



Training by 'bringing to face' yields to best accuracy

Experiment 2 - conclusions

- Can learn more from different scales than from clean, as long as scale 3 is included
- Learning from different scales gives better accuracies when tested on lower scales
- Test on clean much better than test on scales

Scale 0
10% of view

Scale 1
20% of view

Scale 2
30% of view

Scale 3
60% of view

Clean
Image

Image source: collages we made from Caltech 256 database

Conclusions

- With our controlled datasets, we could verify that network learns better from larger scale
- Testing needs to be done on clean images, no matter which scales were used in training
- Training on scales >30% gives more robustness when testing on all scales
- Training on scales <30% hurts accuracy