# ShapeMask: Learning to Segment Novel Objects by Refining Shape Priors

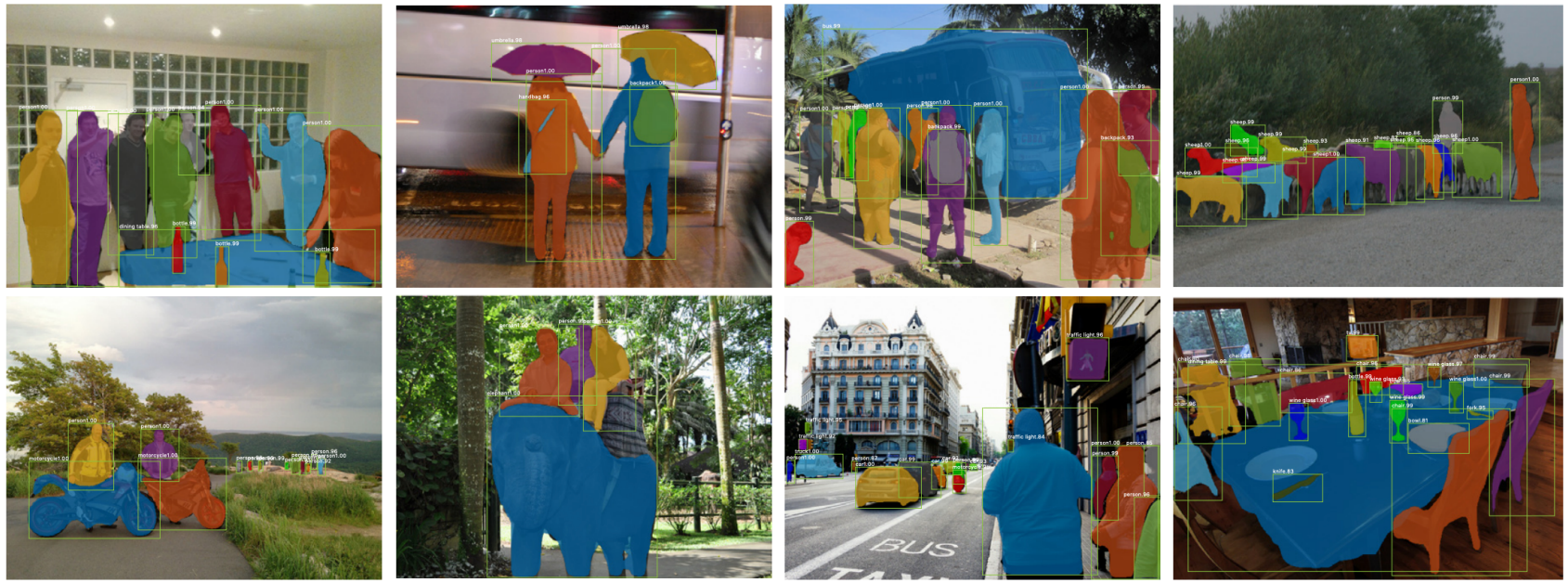
Weicheng Kuo[1], Anelia Angelova[1], Jitendra Malik[2], Tsung-Yi Lin[1]
[1] Google Brain [2] University of California, Berkeley
[1]{weicheng, anelia, tsungyi}@google.com, [2] malik@eecs.berkeley.edu

Organized by: Xi Fang

Date: 04/08/2020

# Background

- Instance Segmentation
  - pixel- level classification of objects and identifying individual objects as separate entities

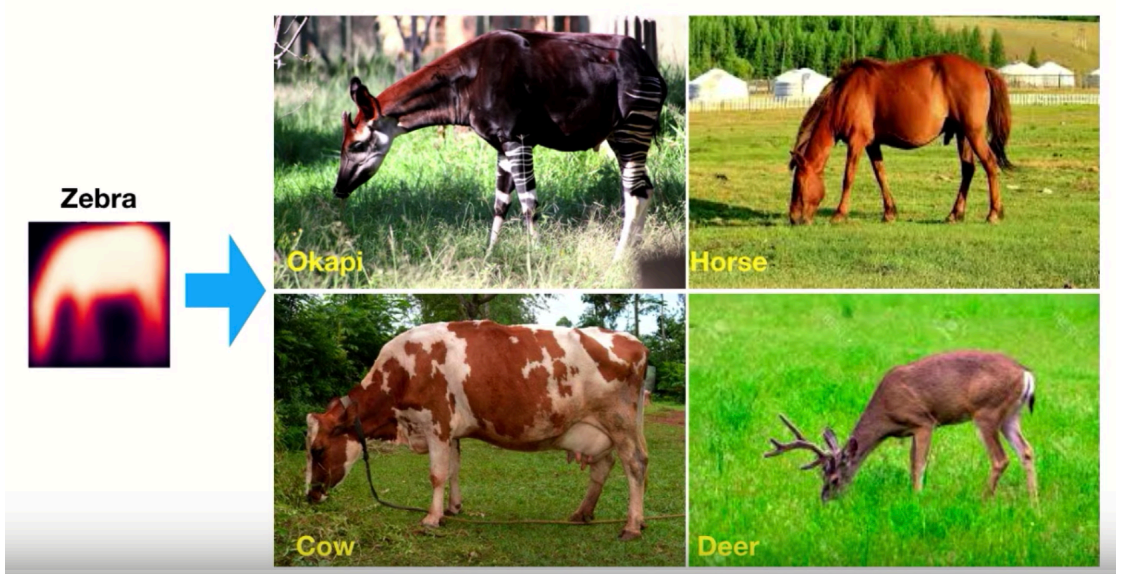


He et al. Mask R-CNN, ICCV 2017

# Motivation

- Requires mask labels for every category
- The real world contains many long-tailed and open-set categories
- Clustering and transfer learning have been used to tackle the problem

- **Shape priors** are common across categories
- Some classical works use shape priors to segmentation



# Contribution

- ShapeMask significantly outperforms the state-of-the-art **transfer learning** approach in the cross-category setup
  - Outperform the state-of-the-art using only 1% of the labeled data.
- ShapeMask is able to segment many **novel object** classes in a robotics environment different from the COCO dataset.
- ShapeMask is competitive with **state-of-the-art** techniques while training multiple times faster and testing at 150-200ms per image.

# Method

- Intermediate representations

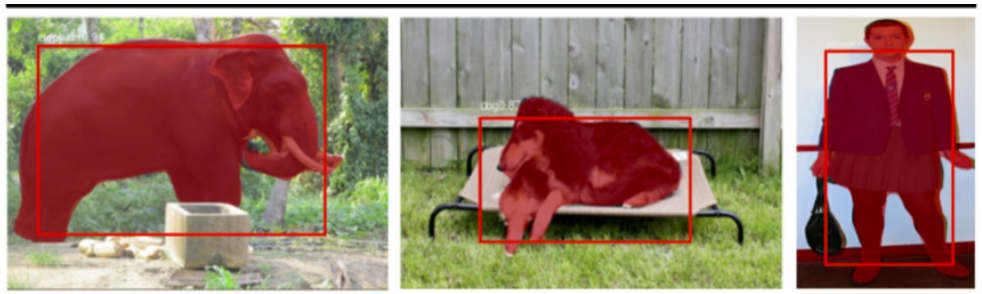
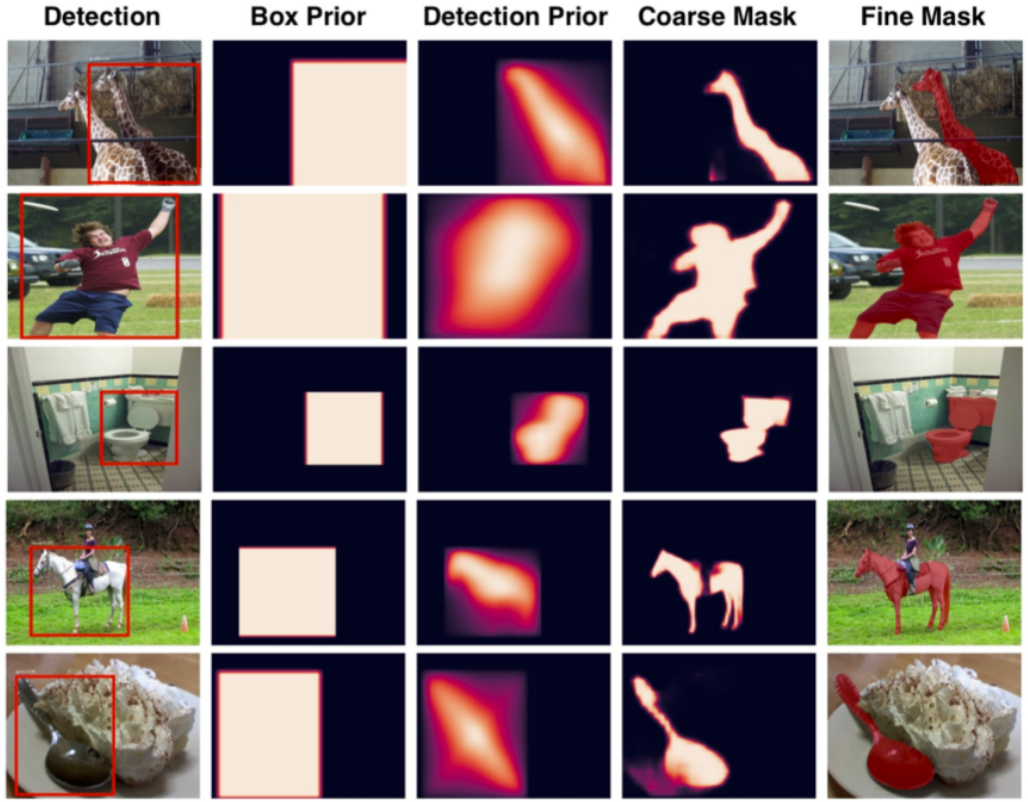
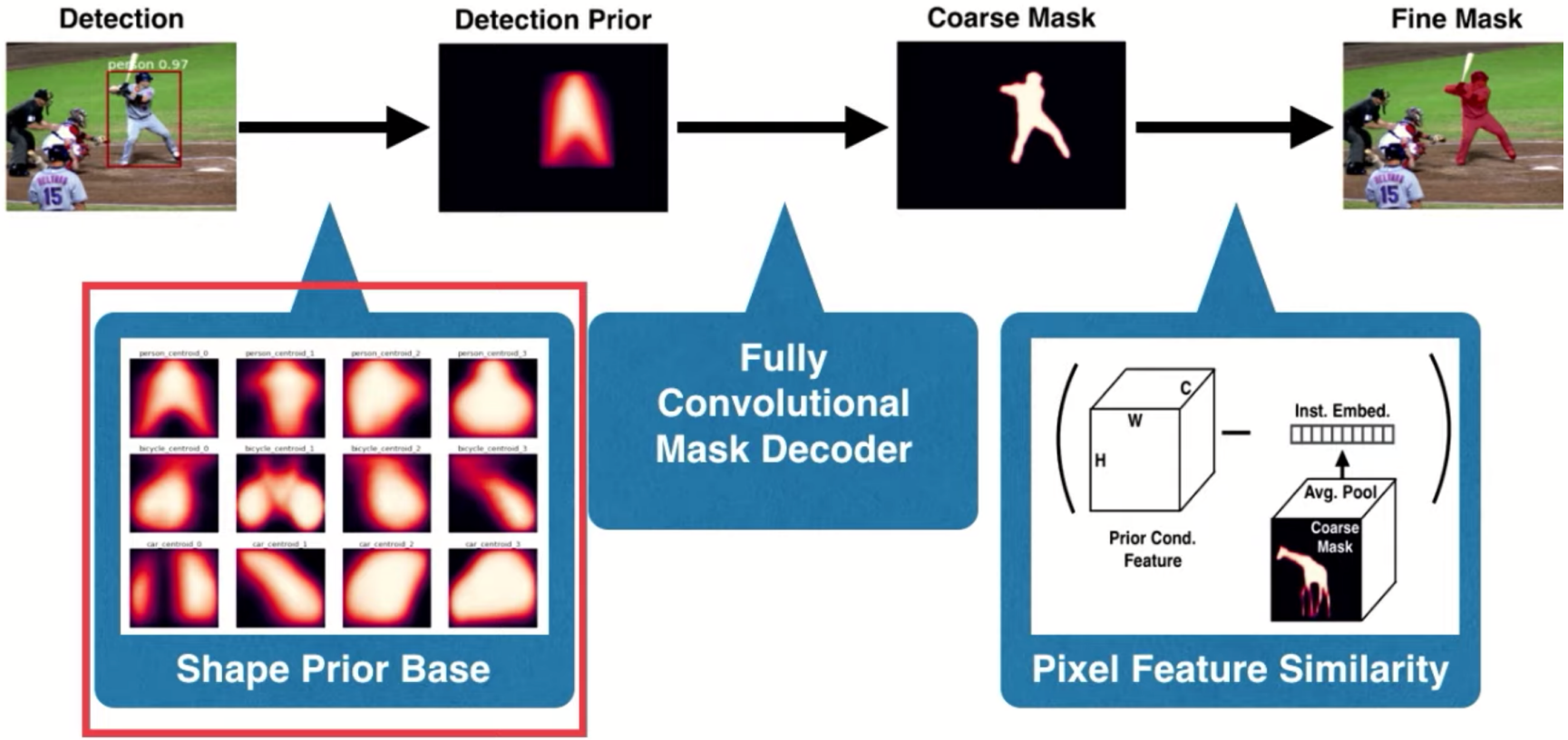


Figure 1: ShapeMask instance segmentation is designed to learn the shape of objects by refining object shape priors. Starting from a bounding box (leftmost column), the shape is progressively refined in our algorithm until reaching the final mask (rightmost column). The bounding box is only needed to approximately localize the object of interest and is not required to be accurate (bottom row).



# Method

- ShapeMask



# Method

- Shape priors
  - Cluster the training masks and take the centroids as shape priors

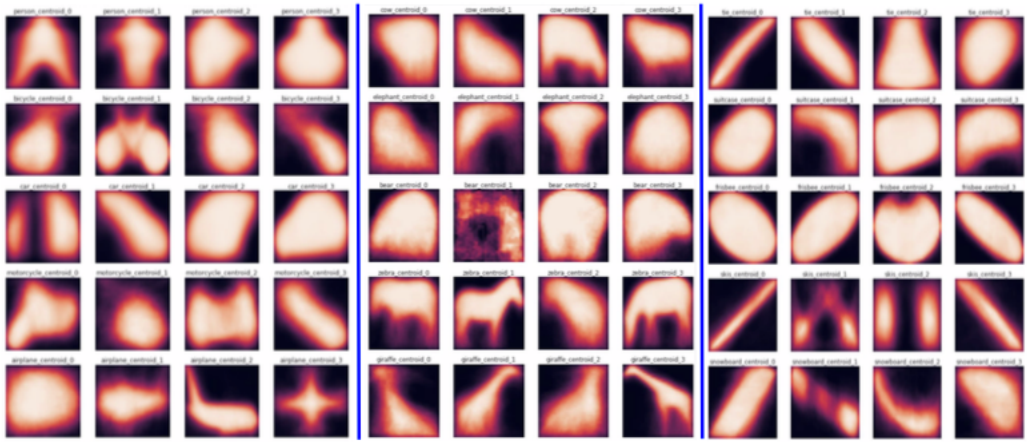


Figure 3: Shape priors obtained by clustering mask labels in the training set. Each prior is a cluster centroid of an object category.

# Method

- Shape priors
  - Cluster the training masks and take the centroids as shape priors

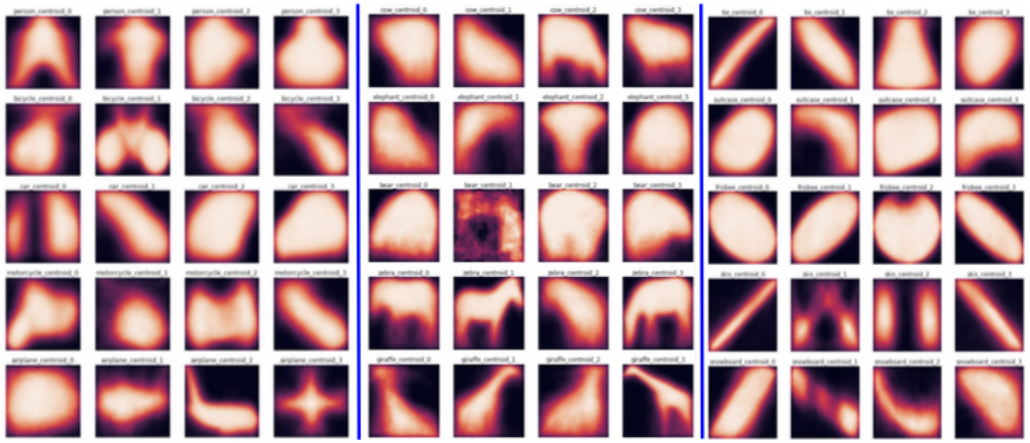


Figure 3: Shape priors obtained by clustering mask labels in the training set. Each prior is a cluster centroid of an object category.

total number of shape priors is C × K

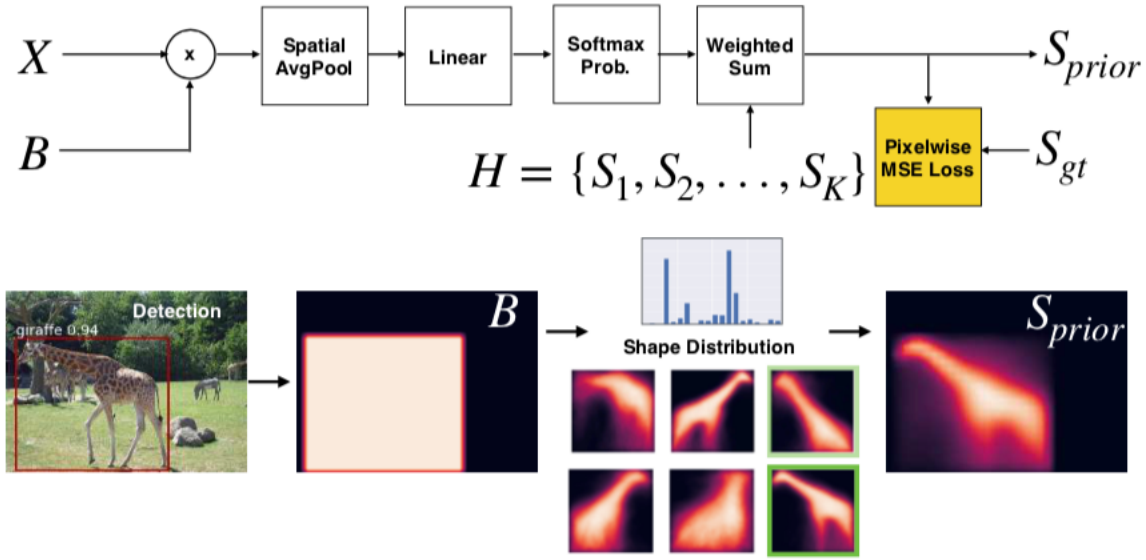
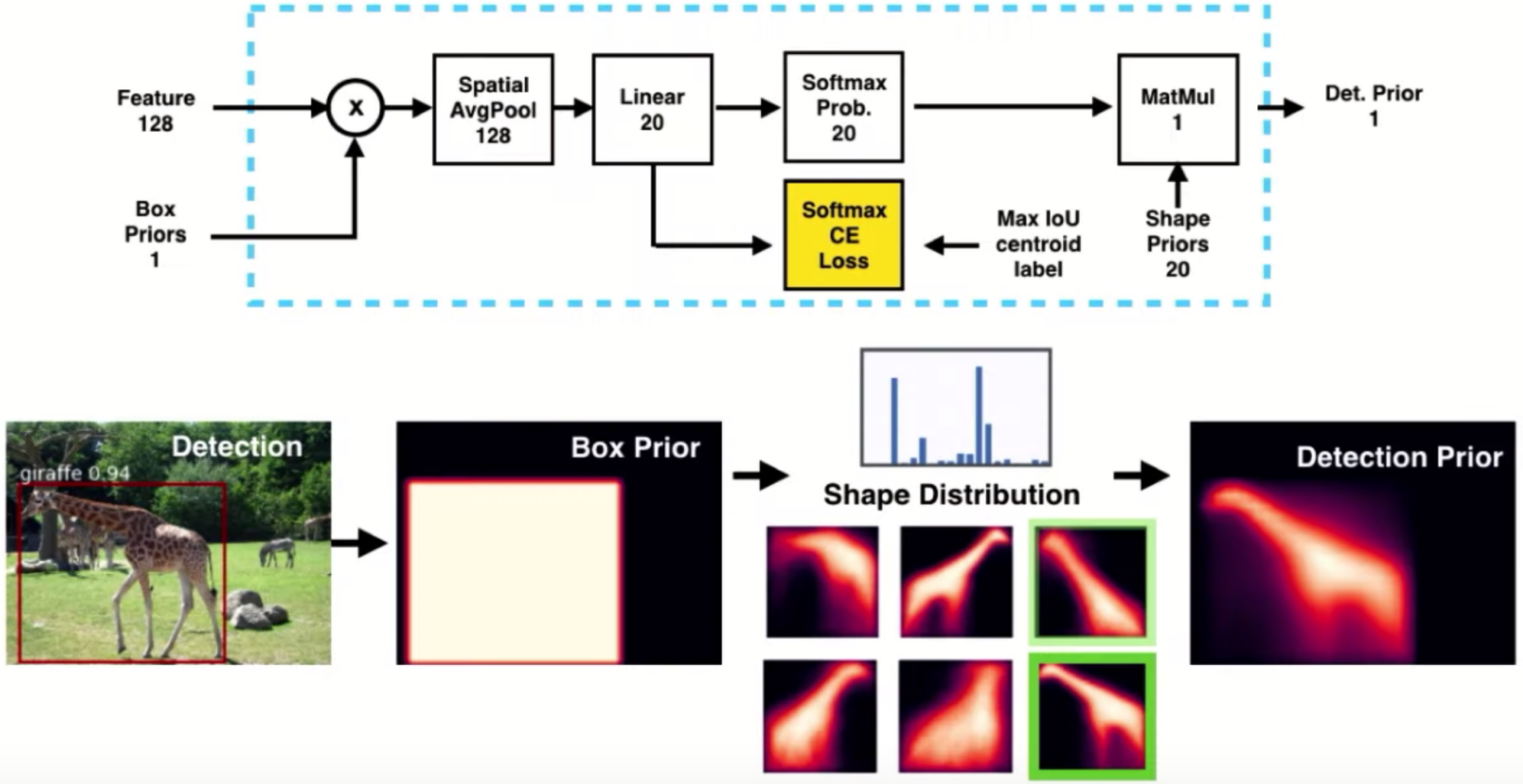



Figure 4: **Shape Estimation.** Given a box detection, we refine the box into an initial estimate of shape $S_{prior}$ by linearly combining prior shapes $S_1, S_2, \ldots, S_k$. Our model learns to predict the shape prior distribution to minimize reconstruction error.

# Method

- Detection Prior Learning
  - Predict the shape prediction over cluster centroids



# Method

- Coarse Mask Prediction
  - Low resolution mask from the detection prior

$$X_{prior} = X + g(S_{prior})$$

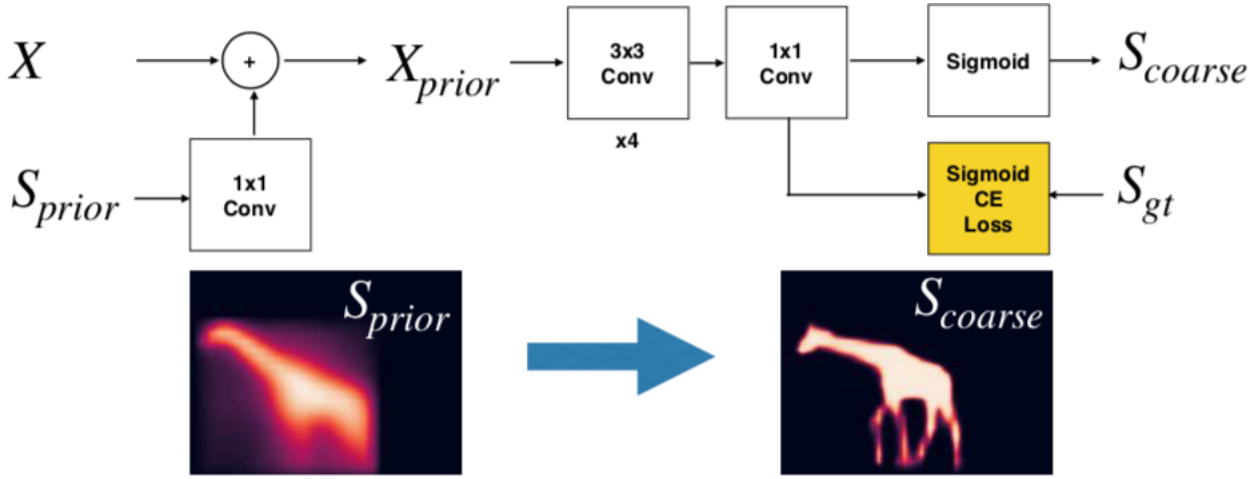
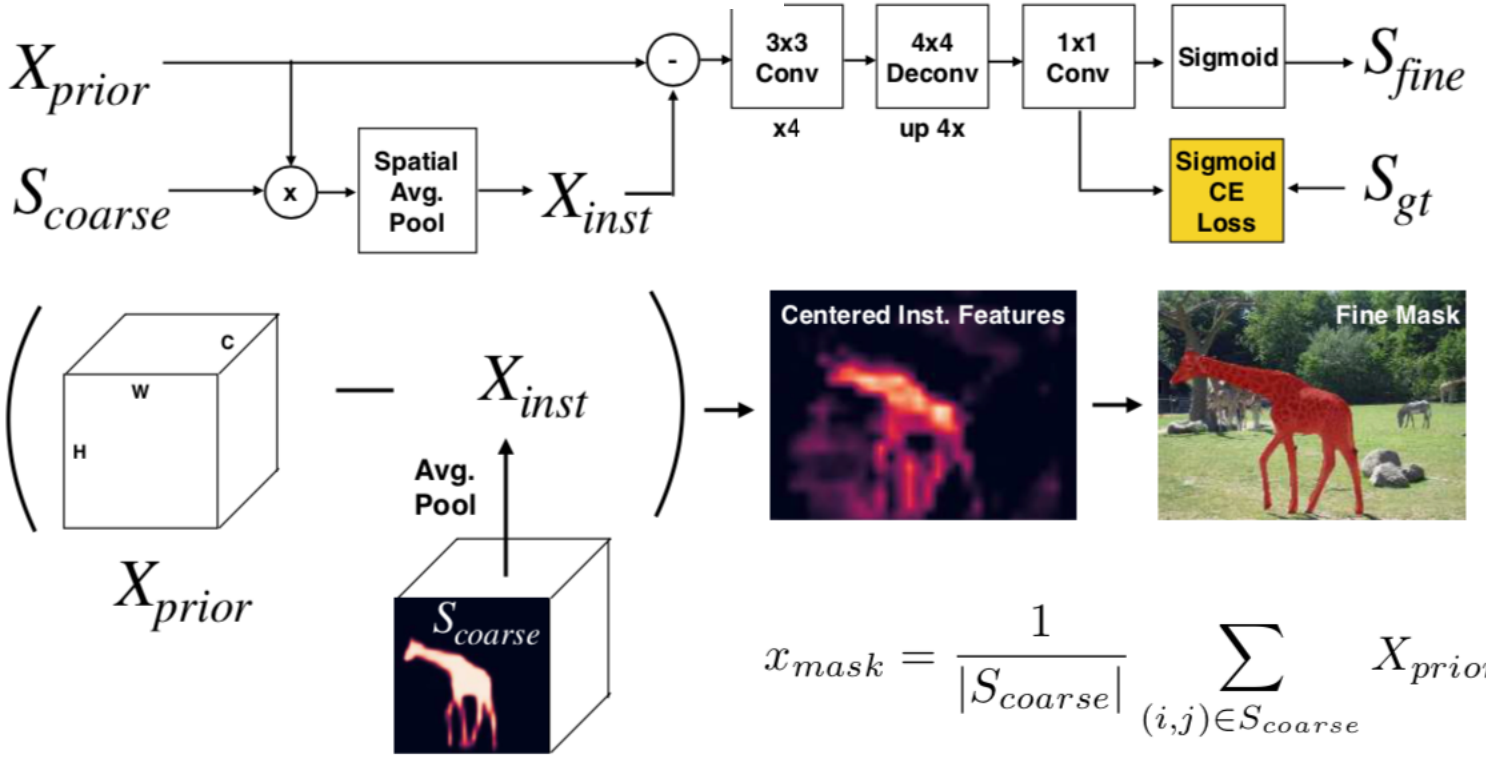



Figure 5: **Coarse Mask Prediction.** We fuse $S_{prior}$ with the image features $X$ to obtain prior conditioned features $X_{prior}$, from which we decode a coarse shape $S_{coarse}$.

# Method

- Fine Mask Prediction

$$X_{inst(i,j)} = X_{prior(i,j)} - x_{mask}$$




$$x_{mask} = \frac{1}{|S_{coarse}|} \sum_{(i,j) \in S_{coarse}} X_{prior(i,j)}$$

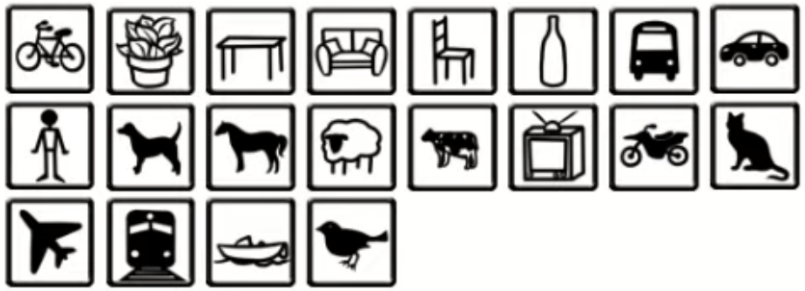
# Experiment

- Generalization to Novel Categories
  - Train on A (box, mask) + B (box). Evaluate on B (mask)

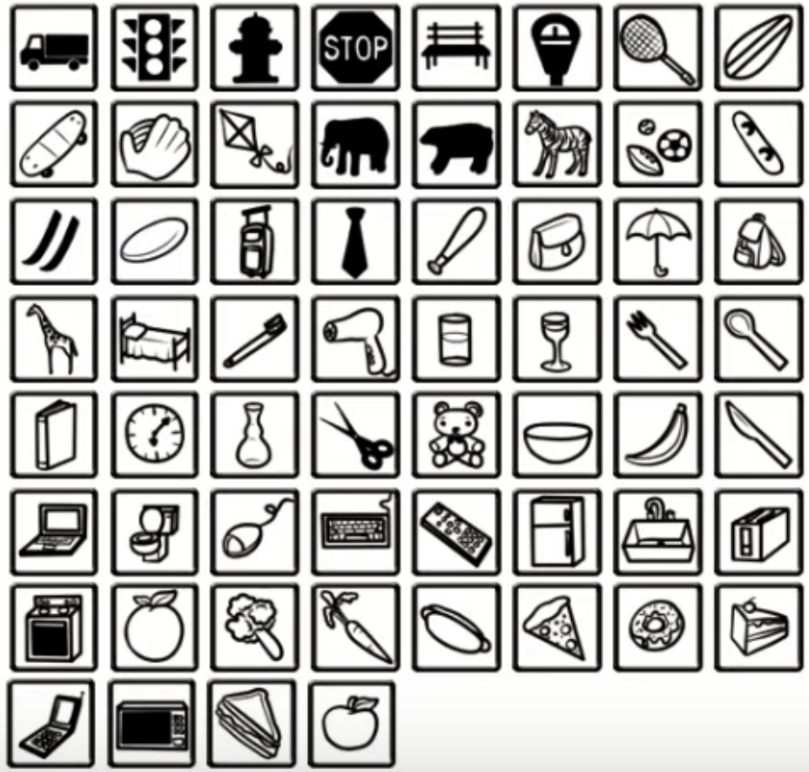
A

B

Set A (20 VOC)



Set B (60 Non-VOC)



# Results

# Generalization to Novel Categories

- Comparison with state-of-the-art - Mask RCNN

| | | voc → non-voc | | | | | | non-voc → voc | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| backbone | method | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
| | Mask R-CNN [21] | 18.5 | 34.8 | 18.1 | 11.3 | 23.4 | 21.7 | 24.7 | 43.5 | 24.9 | 11.4 | 25.7 | 35.1 |
| | Our Mask R-CNN | 21.9 | 39.6 | 21.9 | 16.1 | 29.7 | 24.6 | 27.2 | 39.6 | 27.0 | 16.4 | 31.8 | 35.4 |
| FPN | GrabCut Mask R-CNN [21] | 19.7 | 39.7 | 17.0 | 6.4 | 21.2 | 35.8 | 19.6 | 46.1 | 14.3 | 5.1 | 16.0 | 32.4 |
| | $Mask^X$ R-CNN [21] | 23.8 | 42.9 | 23.5 | 12.7 | 28.1 | 33.5 | 29.5 | 52.4 | 29.7 | 13.4 | 30.2 | 41.0 |
| | Oracle Mask R-CNN [21] | 34.4 | 55.2 | 36.3 | 15.5 | 39.0 | 52.6 | 39.1 | 64.5 | 41.4 | 16.3 | 38.1 | 55.1 |
| | Our Oracle Mask R-CNN | 34.3 | 54.7 | 36.3 | 18.6 | 39.1 | 47.9 | 38.5 | 64.4 | 40.4 | 18.9 | 39.4 | 51.4 |
| FPN | ShapeMask (ours) | **30.2** | **49.3** | **31.5** | **16.1** | **38.2** | **38.4** | **33.3** | **56.9** | **34.3** | **17.1** | **38.1** | **45.4** |
| | Oracle ShapeMask (ours) | 35.0 | 53.9 | 37.5 | 17.3 | 41.0 | 49.0 | 40.9 | 65.1 | 43.4 | 18.5 | 41.9 | 56.6 |
| NAS-FPN [13] | ShapeMask (ours) | **33.2** | **53.1** | **35.0** | **18.3** | **40.2** | **43.3** | **35.7** | **60.3** | **36.6** | **18.3** | **40.5** | **47.3** |
| | Oracle ShapeMask (ours) | 37.6 | 57.7 | 40.2 | 20.1 | 44.4 | 51.1 | 43.1 | 67.9 | 45.8 | 20.1 | 44.3 | 57.8 |

Table 1: Performance of ShapeMask (class-agnostic) on novel categories. At the top, voc → non-voc means "train on masks in voc, test on masks in non-voc", and vice versa. ShapeMask outperforms the state-of-the-art method $Mask^X$ R-CNN [21] by 6.4 AP on voc to non-voc transfer, and 3.8 AP on non-voc to voc transfer using the same ResNet backbone. ShapeMask has smaller gap with the oracle upper-bound than $Mask^X$ R-CNN. By using a stronger feature pyramid from [13], ShapeMask outperforms $Mask^X$ R-CNN by 9.4 and 6.2 AP.

# Visualization

# Generalization with less data

ShapeMask generalizes well down to 1/1000 of the training data.

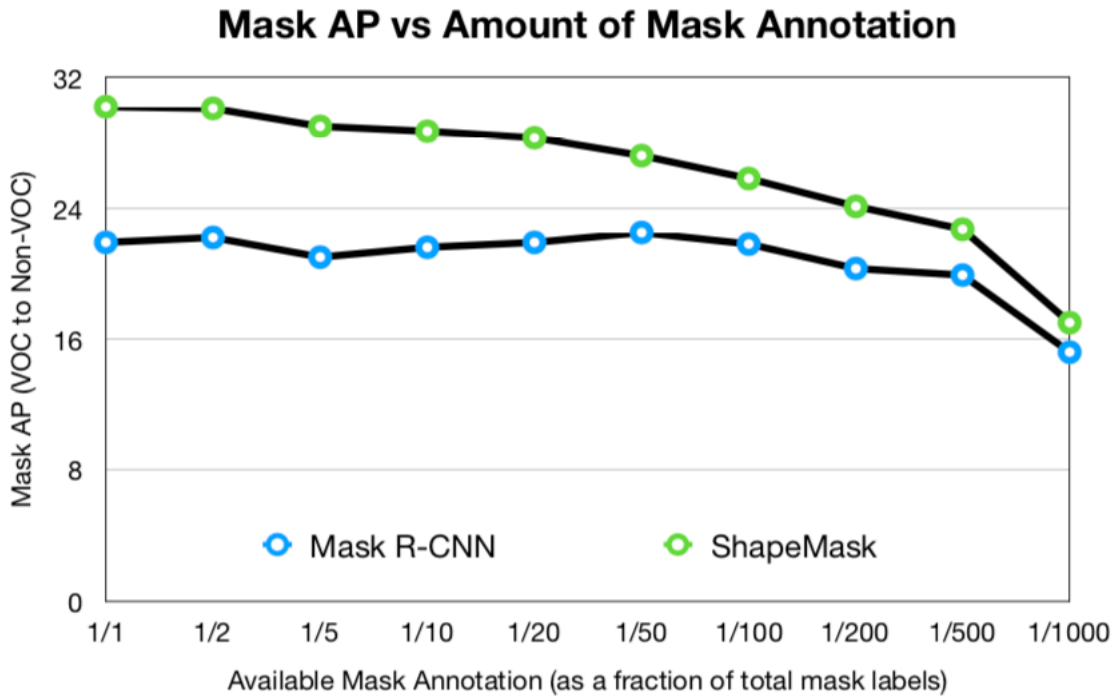



Figure 8: Generalization with less data. ShapeMask generalizes well down to 1/1000 of the training data.

# Generalization to robotics data

- ShapeMask works well on many unseen categories in an office environment (trained on COCO only)

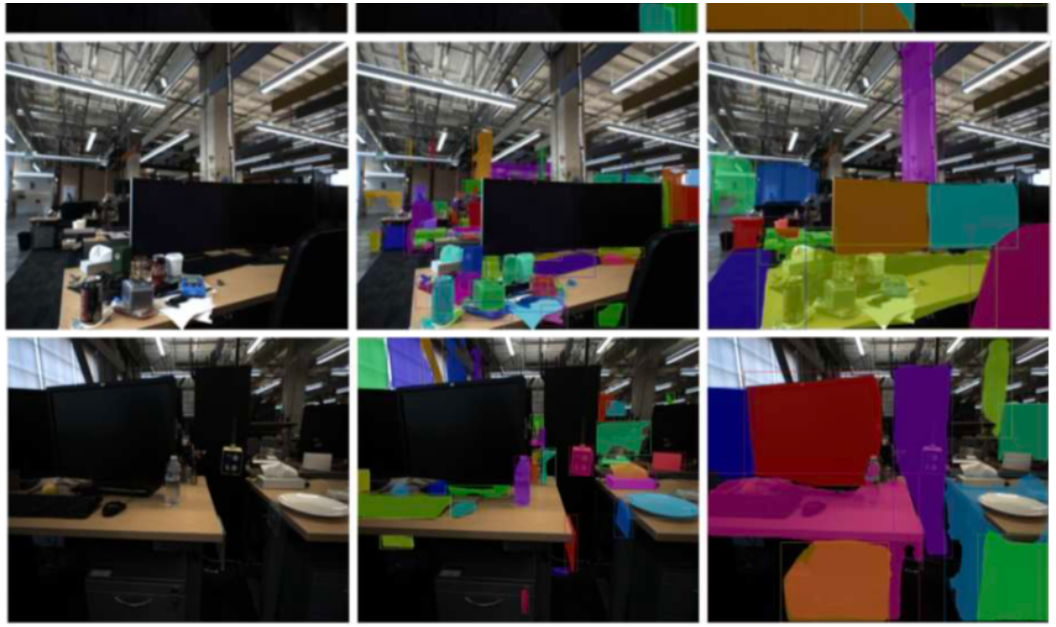


Figure 9: ShapeMask applied for object instance segmentation for robotics grasping. Here the ShapeMask model is trained on the COCO dataset and is not fine-tuned on data from this domain. As seen, it successfully segments the object instances, including novel objects such as a plush

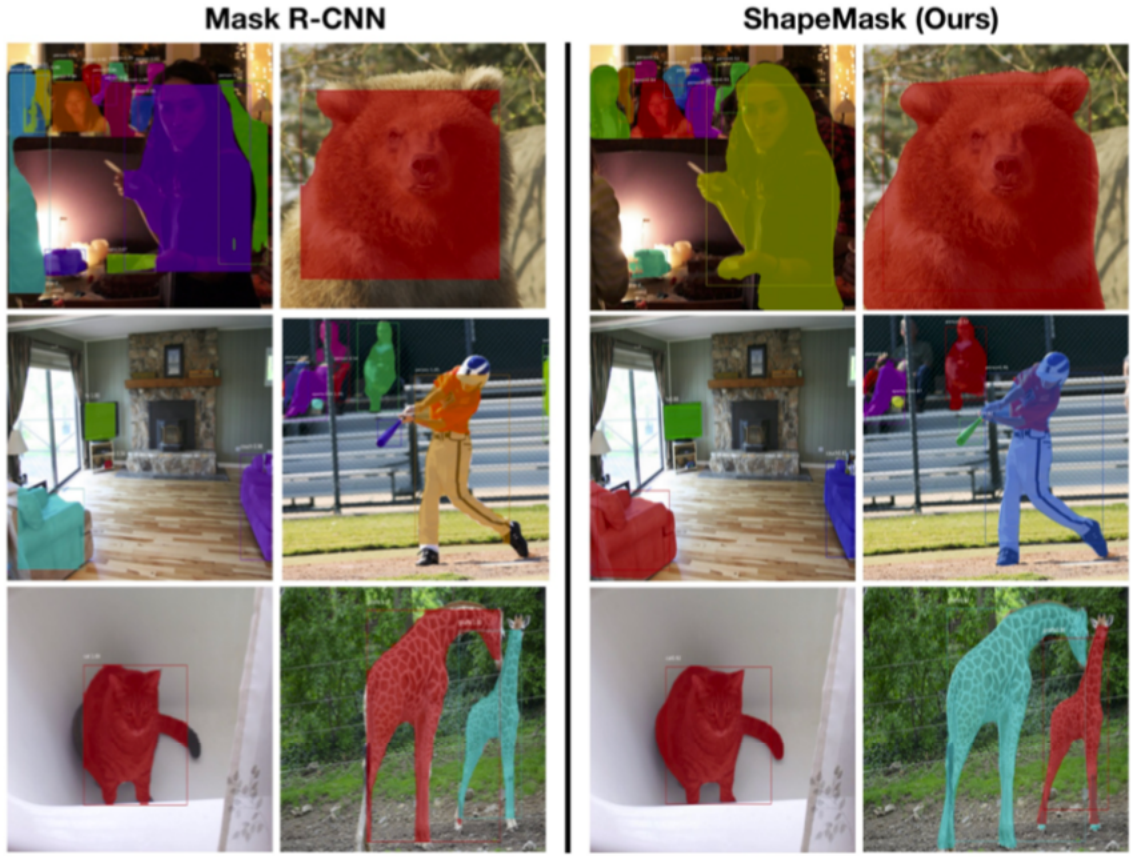
# Fully Supervised Instance Segmentation

| | backbone | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ | Training (hrs) | Inference (X + Y ms) | GPU |
|---|---|---|---|---|---|---|---|---|---|---|
| FCIS+++ [28] +OHEM | ResNet-101-C5-dilate | 33.6 | 54.5 | - | - | - | - | 24 | 240 | K40 |
| Mask R-CNN [19] | ResNet-101-FPN | 35.7 | 58.0 | 37.8 | 15.5 | 38.1 | 52.4 | 44 | 195 + 15 | P100 |
| Detectron Mask R-CNN [15] | ResNet-101-FPN | 36.4 | - | - | - | - | - | 50 | 126 + 17 | P100 |
| ShapeMask (ours) | ResNet-101-FPN | 37.4 | 58.1 | 40.0 | 16.1 | 40.1 | 53.8 | 11* | 125 + 24 | V100 |
| Mask R-CNN [19] | ResNext-101-FPN | 37.1 | 60.0 | 39.4 | 16.9 | 39.9 | 53.5 | - | - | - |
| MaskLab [6] | Dilated ResNet-101 | 37.3 | 59.8 | 39.6 | 19.1 | 40.5 | 50.6 | - | - | - |
| PANet [34] | ResNext-101-PANet | 42.0 | 65.1 | 45.7 | 22.4 | 44.7 | 58.1 | - | - | - |
| ShapeMask (ours) | ResNet-101-NAS-FPN [13] | 40.0 | 61.5 | 43.0 | 18.3 | 43.0 | 57.1 | 25* | 180 + 24 | V100 |

Table 2: ShapeMask Instance Segmentation Performance on COCO. Using the same backbone, ShapeMask outperforms Mask R-CNN by 1.7 AP. With a larger backbone, ShapeMask outperforms Mask R-CNN and MaskLab by 2.9 and 2.7 AP respectively. Compared to PANet, ShapeMask is only 2.0 AP behind without using any techniques reported in [34, 6]. This shows that ShapeMask is competitive in the fully supervised setting. Timings reported on TPUs are marked with star signs. Inference time is reported following the Detectron format: X for GPU time, Y for CPU time. All mask APs are single-model, and are reported on COCO test-dev2017 without test time augmentation except Detectron on val2017 (gray).

# Fully Supervised Instance Segmentation

- Analysis of robust segmentation



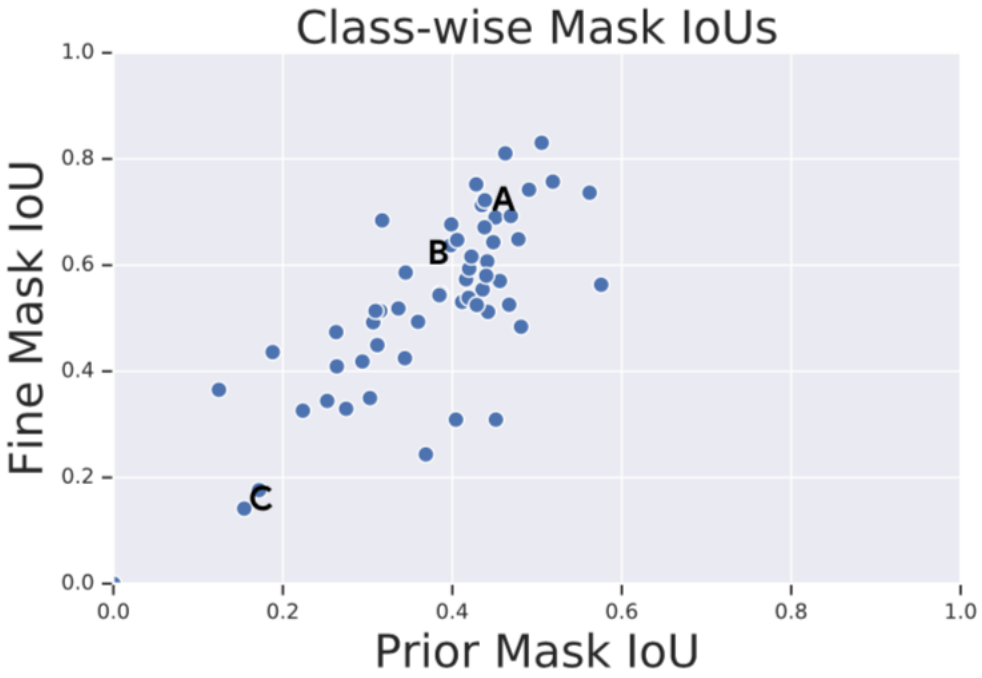
# Ablation Study

- Compare the uniform **box prior** with our learned detection prior, and the direct **mask decoding** with our instance conditioned mask decoding
- COCO val2017 using ResNet-101-FPN

| | | voc → non-voc | | | non-voc → voc | | |
|---|---|---|---|---|---|---|---|
| Shape | Embed. | AP | $AP_{50}$ | $AP_{75}$ | AP | $AP_{50}$ | $AP_{75}$ |
| | | 13.7 | 28.0 | 12.0 | 24.8 | 45.6 | 23.5 |
| | ✓ | 26.2 | 44.6 | 27.1 | 29.4 | 51.7 | 29.0 |
| ✓ | | 26.4 | 44.9 | 27.2 | 30.6 | 53.4 | 30.4 |
| ✓ | ✓ | 30.2 | 49.3 | 31.5 | 33.3 | 56.9 | 34.3 |

Table 4: Ablation results for the partially supervised model.

# Does prior matter?



# Conclusion

- ShapeMask uses shape priors and instance embeddings for better generalization to novel categories.
- ShapeMask significantly outperforms state-of-the- art in the cross categories setup .
- It is robust against inaccurate detections, competitive in the fully supervised setting, and runs efficiently for training and inference