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

1. Overview

This is the supplementary materials to accompany the ShapeMask paper [6]. Here we present additional experiments, visualizations, an onboard demonstration and more implementation details.

2. Fully Supervised ShapeMask

Visualization. Figure 1 visualizes the outputs of the fully supervised ShapeMask. ShapeMask obtains quality contours for large objects (e.g. humans and animals) and can handle thin structures (e.g. legs of chairs, skateboard) and within-category overlaps (e.g. crowds of humans). Results are generated by class-specific ResNet-101-FPN model on COCO validation set.

Object detection results. Instance segmentation algorithms are also evaluated by their ability to provide accurate detections [4]. In addition to the instance segmentation results in Table 2 of the manuscript, we compare ShapeMask with leading object detectors in Table 1 on COCO. With ResNet-101-FPN backbone, our 42.0 AP clearly outperforms RetinaNet and Mask R-CNN, and is among the best reported approaches using the same backbone. Applying a stronger backbone NAS-FPN [3] (as described below), ShapeMask achieves 45.4 AP which is comparable to SNIP[11] and behind PANet by 2.0 AP. Note that ShapeMask does not apply any detection improvement methods proposed in [1, 5, 11, 9]. This shows that ShapeMask can function as a competitive object detector as well.

3. Mask Branch Runtime

We study the performance and the mask branch capacity tradeoff in Table 2. All convolution and deconvolution layers in the mask branch are set to the same number of channels here. We observe that ShapeMask performance degrades minimally as the mask branch capacity decreases dramatically. With 16 channels, the mask branch of ShapeMask maintains a competitive AP of 35.8, slightly better Mask R-CNN, while using 130x fewer parameters and 23x fewer FLOPs and running at 4.6ms. To our knowledge, this is the most **lightweight and yet competitive mask branch** design for instance segmentation.

4. Model Ablation

To understand our system further, we compare the uniform box prior with the learned detection prior, and the direct mask decoding [4] with the instance conditioned mask decoding. Table 3 shows the fully supervised system ablation results on COCO val2017 using ResNet-101-FPN. Using either object shape prior or instance embedding improves from the baseline. Combining both techniques boosts the performance even further. This demonstrates the importance of the key components of our algorithm, namely shape priors and learned embeddings.

5. Applying a different backbone

In ShapeMask framework, the backbone ResNet-101-FPN model is used for efficiency, but can be easily replaced by a stronger backbone to improve the accuracy. More specifically, in the paper we also experiment with [3]. The replacement of this model, as well as others, is very easy to do. More specifically, we replace the FPN connections by NAS-FPN connections which keeps all feature dimensions unchanged and allows us to run the mask branch on the same input shapes as before. We note that this provides gains in accuracy at a very small computational cost (e.g. from 150ms vs 200ms for NAS-FPN) Other strong backbone models such as ResNext-101-FPN [4] can also be used.

6. Onboard Demonstration

Figure 2 visualizes the outputs of class-agnostic ShapeMask running onboard a robot. ShapeMask is able to capture detailed contours of many novel objects. The model is trained on COCO only, so this is an out-of-sample setting. Results are generated by ResNet-101-FPN using inference time optimization platform TensorRT.

Figure 1: Visualization of results of the fully supervised ShapeMask model on the COCO val2017. ShapeMask is able to obtain quality contours for large objects, handle thin structures, and deal with within-category overlaps.

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN [4]	ResNet-101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
RetinaNet [8]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
MaskLab [2]	Dilated ResNet 101	41.9	62.6	46.0	23.8	45.5	54.2
Cascade R-CNN [1]	ResNet-101	42.8	62.1	46.3	23.7	45.5	55.2
SNIP [11]	DPN-98	45.7	67.3	51.1	29.3	48.8	57.1
PANet [9]	Path-Agg. ResNext-101	47.4	67.2	51.8	30.1	51.7	60.0
ShapeMask (ours)	ResNet-101-FPN	42.0	61.2	45.7	24.3	45.2	53.1
ShapeMask (ours)	ResNet-101-NAS-FPN [3]	45.4	64.2	49.2	27.2	49.0	56.9

Table 1: Object Detection Box AP on COCO test-dev2017. With ResNet-101-FPN backbone, ShapeMask outperforms RetinaNet and Mask R-CNN, and is among the best reported approaches using the same backbone. With a larger backbone, ShapeMask achieves comparable performance to SNIP and trails PANet by 2 points without using any techniques from [1, 5, 11]. All entries are single model results and without test time augmentation.

Model	# of Chns.	AP	Params. (M)	FLOPs (M)	Time (ms)
Mask R-CNN [4]	256	35.4	2.64	530	-
ShapeMask (ours)	128	37.0	1.44	1480	29.1
ShapeMask (ours)	64	36.7	0.36	370	14.0
ShapeMask (ours)	32	36.6	0.09	93	7.0
ShapeMask (ours)	16	35.8	0.02	23	4.6

Table 2: Performance vs. mask branch model capacity. The performance decreases only slightly with a dramatic decrease in the model capacity of the mask branch. With only 16 channels, ShapeMask model achieves 0.4 AP higher than Mask R-CNN with 130x fewer parameters and 23x fewer FLOPs. Timing is measured on the mask branch only.

Shape	Embed.	AP	AP_{50}	AP_{75}
		35.5	56.5	37.9
	✓	36.7	57.3	38.9
✓		36.9	57.3	39.6
✓	✓	37.2	57.6	39.6

Table 3: Ablation results for the fully supervised model.

7. Implementation Details

One-stage detection: We adopt RetinaNet[1] [8] to generate bounding box detections for ShapeMask. RetinaNet is a one-stage detector with a simple system design and competitive performance. We use an image size of 1024 x 1024 and multiscale training with image scale from 0.8 to 1.2. Note that other detection methods such as Faster R-CNN [10] can also be used with ShapeMask.

RoI features: We use the feature pyramid [7] with levels P_3 to P_5 to process RoI in different scales for scale normalization. Given a bounding box, we assign the box to feature

[1] https://github.com/tensorflow/tpu/tree/master/models/official/retinanet

Figure 2: Visualization of ShapeMask running **onboard a robot** for instance segmentation of table-top objects for the purposes of grasping. ShapeMask is able to segment many novel objects well, e.g. paper, headphone. This is an out-of-sample setting because the model is trained on COCO only, but tested on indoor office scenes.

level:

$$k = m - \lfloor \log_2 \frac{L}{\max(box_h, box_w)} \rfloor, \quad (1)$$

where L is the image size (e.g., 1024) and m is the highest feature level (e.g., 5). If k is less than the minimum level (e.g. 3), the box is assigned to minimum level. At the assigned level, we take a $c \times c$ feature patch centered on the box. We choose $c = L/2^m$ to make sure the entire object instance always lies inside the patch by the feature pyramid design [7]. The feature dimension is then reduced from 256 to 128 by 1x1 convolution. We apply ShapeMask algorithm on this feature patch X to predict the instance mask (the same X in figure 5 of the paper). This is a simple slicing operation, but we note that the "crop and resize" operation [4, 2] could also work.

Training with jittered groundtruths: Unlike [4, 2] which sample masks from the object proposals, we directly sample 8 groundtruth masks and their associated boxes per image for training. This removes the need of object proposal stage and enables one-stage training for mask prediction. The sampled groundtruth boxes are jittered by gaussian noise to better mimic the imperfect boxes produced by the model during inference time. To be precise, the new box center $(x'_c, y'_c) = (x_c + \delta_x w, y_c + \delta_y h)$, and the new box size $(w', h') = (e^{\delta_w} w, e^{\delta_h} h)$, where (x, y, w, h), (x', y', w', h') are the noiseless/jittered groundtruth boxes respectively, and δs is gaussian noise $\sim N(\mu = 0, \sigma = 0.1)$. We represent these boxes by uniform box priors (see B in figure 4 of the paper) for training the mask branch. Jittering is essential to help ShapeMask learn to be robust against noisy detections at test time.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: delving into high quality object detection. *arXiv preprint arXiv:1712.00726*, 2017.

[2] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. *arXiv preprint arXiv:1712.04837*, 2017.

[3] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V. Le. NAS-FPN: learning scalable feature pyramid architecture for object detection. In *CVPR*, 2019.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[5] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. *arXiv preprint arXiv:1807.11590*, 1, 2018.

[6] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. *arXiv preprint arXiv:1904.03239*, 2019.

[7] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017.

[8] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[9] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8759–8768, 2018.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[11] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection–snip. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3578–3587, 2018.