# YOLACT
# Real-time Instance Segmentation

Daniel Bolya Chong Zhou Fanyi Xiao Yong Jae Lee
University of California, Davis
{dbolya, cczhou, fyxiao, yongjaelee}@ucdavis.edu

## Abstract

*We present a simple, fully-convolutional model for real-time instance segmentation that achieves 29.8 mAP on MS COCO at 33.5 fps evaluated on a single Titan Xp, which is significantly faster than any previous competitive approach. Moreover, we obtain this result after training on* ***only one GPU****. We accomplish this by breaking instance segmentation into two parallel subtasks: (1) generating a set of prototype masks and (2) predicting per-instance mask coefficients. Then we produce instance masks by linearly combining the prototypes with the mask coefficients. We find that because this process doesn't depend on repooling, this approach produces very high-quality masks and exhibits temporal stability for free. Furthermore, we analyze the emergent behavior of our prototypes and show they learn to localize instances on their own in a translation variant manner, despite being fully-convolutional. Finally, we also propose Fast NMS, a drop-in 12 ms faster replacement for standard NMS that only has a marginal performance penalty.*

## 1. Introduction

*"Boxes are stupid anyway though, I'm probably a true believer in masks except I can't get YOLO to learn them."*
– Joseph Redmon, YOLOv3 [36]

What would it take to create a real-time instance segmentation algorithm? Over the past few years, the vision community has made great strides in instance segmentation, in part by drawing on powerful parallels from the well-established domain of object detection. State-of-the-art approaches to instance segmentation like Mask R-CNN [18] and FCIS [24] directly build off of advances in object detection like Faster R-CNN [37] and R-FCN [8]. Yet, these methods focus primarily on performance over speed, leaving the scene devoid of instance segmentation parallels to real-time object detectors like SSD [30] and YOLO [35, 36]. In this work, our goal is to fill that gap with a fast, one-stage instance segmentation model in the same way that SSD and YOLO fill that gap for object detection.

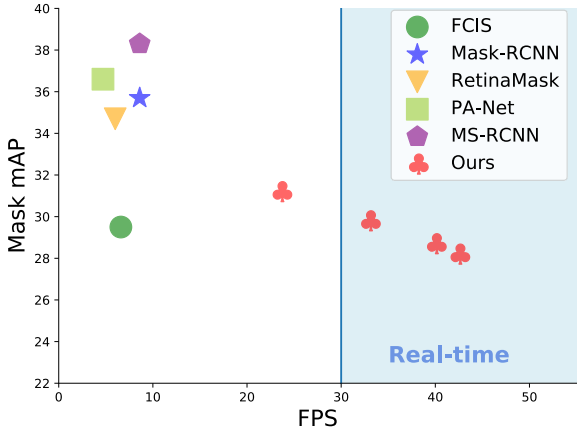


Figure 1: Speed-performance trade-off for various instance segmentation methods on COCO. To our knowledge, ours is the first *real-time* (above 30 FPS) approach with around 30 mask mAP on COCO `test-dev`.

However, instance segmentation is hard—much harder than object detection. One-stage object detectors like SSD and YOLO are able to speed up existing two-stage detectors like Faster R-CNN by simply removing the second stage and making up for the lost performance in other ways. The same approach is not easily extendable, however, to instance segmentation. State-of-the-art two-stage instance segmentation methods depend heavily on *feature localization* to produce masks. That is, these methods "repool" features in some bounding box region (e.g., via RoI-pool/align), and then feed these now localized features to their mask predictor. This approach is inherently sequential and is therefore difficult to accelerate. One-stage methods that perform these steps in parallel like FCIS do exist, but they require significant amounts of post-processing after localization, and thus are still far from real-time.

To address these issues, we propose YOLACT[1], a real-time instance segmentation framework that forgoes an explicit localization step. Instead, YOLACT breaks up instance segmentation into two parallel tasks: (1) generat-

[1] **Y**ou **O**nly **L**ook **A**t **C**oefficien**T**s

ing a dictionary of non-local *prototype masks over the entire image*, and (2) predicting a set of *linear combination coefficients per instance*. Then producing a full-image instance segmentation from these two components is simple: for each instance, linearly combine the prototypes using the corresponding predicted coefficients and then crop with a predicted bounding box. We show that by segmenting in this manner, *the network learns how to localize instance masks on its own*, where visually, spatially, and semantically similar instances appear different in the prototypes.

Moreover, since the number of prototype masks is independent of the number of categories (e.g., there can be more categories than prototypes), YOLACT learns a distributed representation in which each instance is segmented with a combination of prototypes that are shared across categories. This distributed representation leads to interesting emergent behavior in the prototype space: some prototypes spatially partition the image, some localize instances, some detect instance contours, some encode position-sensitive directional maps (similar to those obtained by hard-coding a position-sensitive module in FCIS [24]), and most do a combination of these tasks (see Figure 5).

This approach also has several practical advantages. First and foremost, it's fast: because of its parallel structure and extremely lightweight assembly process, YOLACT adds only a marginal amount of computational overhead to a one-stage backbone detector, making it easy to reach 30 fps even when using ResNet-101 [19]; in fact, *the entire mask branch takes only ∼5 ms to evaluate*. Second, masks are high-quality: since the masks use the full extent of the image space without any loss of quality from repooling, our masks for large objects are significantly higher quality than those of other methods (see Figure 7). Finally, it's general: the idea of generating prototypes and mask coefficients could be added to almost any modern object detector.

Our main contribution is the first real-time ($> 30$ fps) instance segmentation algorithm with competitive results on the challenging MS COCO dataset [28] (see Figure 1). In addition, we analyze the emergent behavior of YOLACT's prototypes and provide experiments to study the speed vs. performance trade-offs obtained with different backbone architectures, numbers of prototypes, and image resolutions. We also provide a novel Fast NMS approach that is 12ms faster than traditional NMS with a negligible performance penalty. The code for YOLACT is available at https://github.com/dbolya/yolact.

## 2. Related Work

**Instance Segmentation** Given its importance, a lot of research effort has been made to push instance segmentation *accuracy*. Mask-RCNN [18] is a representative two-stage instance segmentation approach that first generates candidate region-of-interests (ROIs) and then classifies and segments those ROIs in the second stage. Follow-up works try to improve its accuracy by e.g., enriching the FPN features [29] or addressing the incompatibility between a mask's confidence score and its localization accuracy [20]. These two-stage methods require re-pooling features for each ROI and processing them with subsequent computations, which make them unable to obtain real-time speeds (30 fps) even when decreasing image size (see Table 2c).

One-stage instance segmentation methods generate position sensitive maps that are assembled into final masks with position-sensitive pooling [6, 24] or combine semantic segmentation logits and direction prediction logits [4]. Though conceptually faster than two-stage methods, they still require repooling or other non-trivial computations (e.g., mask voting). This severely limits their speed, placing them far from real-time. In contrast, our assembly step is much more lightweight (only a linear combination) and can be implemented as one GPU-accelerated matrix-matrix multiplication, making our approach very fast.

Finally, some methods first perform semantic segmentation followed by boundary detection [22], pixel clustering [3, 25], or learn an embedding to form instance masks [32, 17, 9, 13]. Again, these methods have multiple stages and/or involve expensive clustering procedures, which limits their viability for real-time applications.

**Real-time Instance Segmentation** While real-time object detection [30, 34, 35, 36], and semantic segmentation [2, 41, 33, 11, 47] methods exist, few works have focused on real-time instance segmentation. Straight to Shapes [21] and Box2Pix [42] can perform instance segmentation in real-time (30 fps on Pascal SBD 2012 [12, 16] for Straight to Shapes, and 10.9 fps on Cityscapes [5] and 35 fps on KITTI [15] for Box2Pix), but their accuracies are far from that of modern baselines. In fact, Mask R-CNN [18] remains one of the fastest instance segmentation methods on semantically challenging datasets like COCO [28] (13.5 fps on $550^2$ px images; see Table 2c).

**Prototypes** Learning prototypes (aka vocabulary or codebook) has been extensively explored in computer vision. Classical representations include textons [23] and visual words [40], with advances made via sparsity and locality priors [44, 43, 46]. Others have designed prototypes for object detection [1, 45, 38]. Though related, these works use prototypes to represent features, whereas we use them to assemble masks for instance segmentation. Moreover, we learn prototypes that are specific to each image, rather than global prototypes shared across the entire dataset.

## 3. YOLACT

Our goal is to add a mask branch to an existing one-stage object detection model in the same vein as Mask R-CNN [18] does to Faster R-CNN [37], but without an explicit fea-

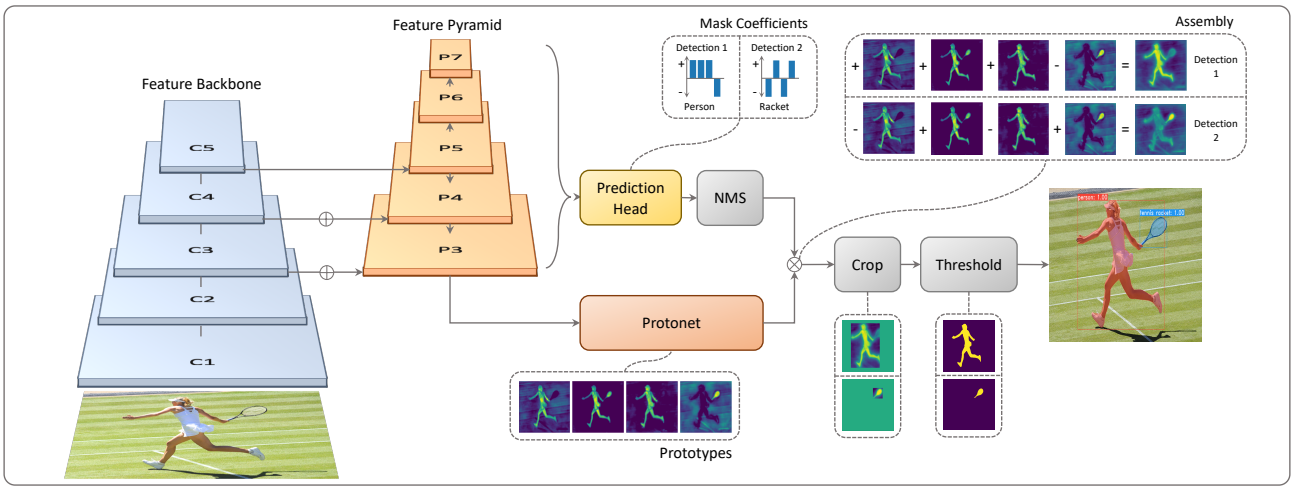


Figure 2: **YOLACT Architecture** Blue/yellow indicates low/high values in the prototypes, gray nodes indicate functions that are not trained, and $k = 4$ in this example. We base this architecture off of RetinaNet [27] using ResNet-101 + FPN.

ture localization step (e.g., feature repooling). To do this, we break up the complex task of instance segmentation into two simpler, parallel tasks that can be assembled to form the final masks. The first branch uses an FCN [31] to produce a set of image-sized "prototype masks" that do not depend on any one instance. The second adds an extra head to the object detection branch to predict a vector of "mask coefficients" for each anchor that encode an instance's representation in the prototype space. Finally, for each instance that survives NMS, we construct a mask for that instance by linearly combining the work of these two branches.

**Rationale** We perform instance segmentation in this way primarily because masks are spatially coherent; i.e., pixels close to each other are likely to be part of the same instance. While a convolutional (*conv*) layer naturally takes advantage of this coherence, a fully-connected (*fc*) layer does not. That poses a problem, since one-stage object detectors produce class and box coefficients for each anchor as an output of an *fc* layer.[2] Two stage approaches like Mask R-CNN get around this problem by using a localization step (e.g., RoIAlign), which preserves the spatial coherence of the features while also allowing the mask to be a *conv* layer output. However, doing so requires a significant portion of the model to wait for a first-stage RPN to propose localization candidates, inducing a significant speed penalty.

Thus, we break the problem into two parallel parts, making use of *fc* layers, which are good at producing semantic vectors, and *conv* layers, which are good at producing spatially coherent masks, to produce the "mask coefficients" and "prototype masks", respectively. Then, because prototypes and mask coefficients can be computed independently, the computational overhead over that of the backbone detector comes mostly from the assembly step, which can be implemented as a single matrix multiplication. In this way, we can maintain spatial coherence in the feature space while still being one-stage and *fast*.

[2]To show that this is an issue, we develop an "*fc*-mask" model that produces masks for each anchor as the reshaped output of an *fc* layer. As our experiments in Table 2c show, simply adding masks to a one-stage model as *fc* outputs only obtains 20.7 mAP and is thus very much insufficient.

### 3.1. Prototype Generation

The prototype generation branch (protonet) predicts a set of $k$ prototype masks for the entire image. We implement protonet as an FCN whose last layer has $k$ channels (one for each prototype) and attach it to a backbone feature layer (see Figure 3 for an illustration). While this formulation is similar to standard semantic segmentation, it differs in that we exhibit no explicit loss on the prototypes. Instead, all supervision for these prototypes comes from the final mask loss after assembly.

We note two important design choices: taking protonet from deeper backbone features produces more robust masks, and higher resolution prototypes result in both higher quality masks and better performance on smaller objects. Thus, we use FPN [26] because its largest feature layers ($P_3$ in our case; see Figure 2) are the deepest. Then, we upsample it to one fourth the dimensions of the input image to increase performance on small objects.

Finally, we find it important for the protonet's output to be unbounded, as this allows the network to produce large, overpowering activations for prototypes it is very confident about (e.g., obvious background). Thus, we have the option of following protonet with either a `ReLU` or no nonlinearity. We choose `ReLU` for more interpretable prototypes.

### 3.2. Mask Coefficients

Typical anchor-based object detectors have two branches in their prediction heads: one branch to predict $c$ class confidences, and the other to predict $4$ bounding box regres-

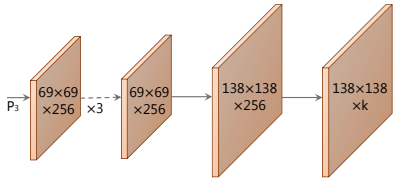


Figure 3: **Protonet Architecture** The labels denote feature size and channels for an image size of $550 \times 550$. Arrows indicate $3 \times 3$ *conv* layers, except for the final *conv* which is $1 \times 1$. The increase in size is an upsample followed by a *conv*. Inspired by the mask branch in [18].

sors. For mask coefficient prediction, we simply add a third branch in parallel that predicts $k$ mask coefficients, one corresponding to each prototype. Thus, instead of producing $4 + c$ coefficients per anchor, we produce $4 + c + k$.

Then for nonlinearity, we find it important to be able to subtract out prototypes from the final mask. Thus, we apply `tanh` to the $k$ mask coefficients, which produces more stable outputs over no nonlinearity. The relevance of this design choice is apparent in Figure 2, as neither mask would be constructable without allowing for subtraction.

### 3.3. Mask Assembly

To produce instance masks, we combine the work of the prototype branch and mask coefficient branch, using a linear combination of the former with the latter as coefficients. We then follow this by a sigmoid nonlinearity to produce the final masks. These operations can be implemented efficiently using a single matrix multiplication and sigmoid:

$$M = \sigma(PC^T) \tag{1}$$

where $P$ is an $h \times w \times k$ matrix of prototype masks and $C$ is a $n \times k$ matrix of mask coefficients for $n$ instances surviving NMS and score thresholding. Other, more complicated combination steps are possible; however, we keep it simple (and fast) with a basic linear combination.

**Losses** We use three losses to train our model: classification loss $L_{cls}$, box regression loss $L_{box}$ and mask loss $L_{mask}$ with the weights 1, 1.5, and 6.125 respectively. Both $L_{cls}$ and $L_{box}$ are defined in the same way as in [30]. Then to compute mask loss, we simply take the pixel-wise binary cross entropy between assembled masks $M$ and the ground truth masks $M_{gt}$: $L_{mask} = \text{BCE}(M, M_{gt})$.

**Cropping Masks** We crop the final masks with the predicted bounding box during evaluation. During training, we instead crop with the ground truth bounding box, and divide $L_{mask}$ by the ground truth bounding box area to preserve small objects in the prototypes.

### 3.4. Emergent Behavior

Our approach might seem surprising, as the general consensus around instance segmentation is that because FCNs

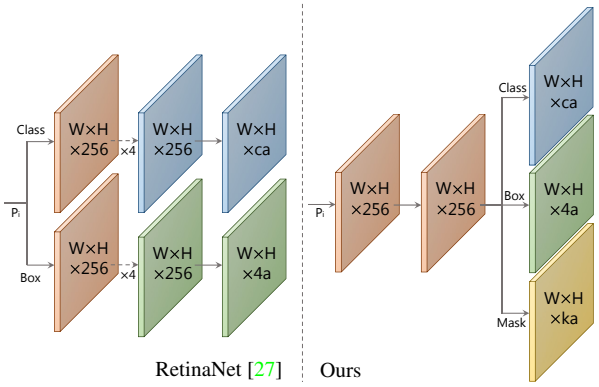


Figure 4: **Head Architecture** We use a shallower prediction head than RetinaNet [27] and add a mask coefficient branch. This is for $c$ classes, $a$ anchors for feature layer $P_i$, and $k$ prototypes. See Figure 3 for a key.

are translation invariant, the task needs translation variance added back in [24]. Thus methods like FCIS [24] and Mask R-CNN [18] try to explicitly add translation variance, whether it be by directional maps and position-sensitive repooling, or by putting the mask branch in the second stage so it does not have to deal with localizing instances. In our method, the only translation variance we add is to crop the final mask with the predicted bounding box. However, we find that our method also works without cropping for medium and large objects, so this is not a result of cropping. Instead, YOLACT *learns how to localize instances on its own* via different activations in its prototypes.

To see how this is possible, first note that the prototype activations for the solid red image (image a) in Figure 5 are actually not possible in an FCN without padding. Because a convolution outputs to a single pixel, if its input everywhere in the image is the same, the result everywhere in the *conv* output will be the same. On the other hand, the consistent rim of padding in modern FCNs like ResNet gives the network the ability to tell how far away from the image's edge a pixel is. Conceptually, one way it could accomplish this is to have multiple layers in sequence spread the padded 0's out from the edge toward the center (e.g., with a kernel like $[1, 0]$). This means ResNet, for instance, *is inherently translation variant*, and our method makes heavy use of that property (images b and c exhibit clear translation variance).

We observe many prototypes to activate on certain "partitions" of the image. That is, they only activate on objects on one side of an implicitly learned boundary. In Figure 5, prototypes 1-3 are such examples. By combining these partition maps, the network can distinguish between different (even overlapping) instances of the same semantic class; e.g., in image d, the green umbrella can be separated from the red one by subtracting prototype 3 from prototype 2.

Furthermore, being learned objects, prototypes are compressible. That is, if protonet combines the functionality of

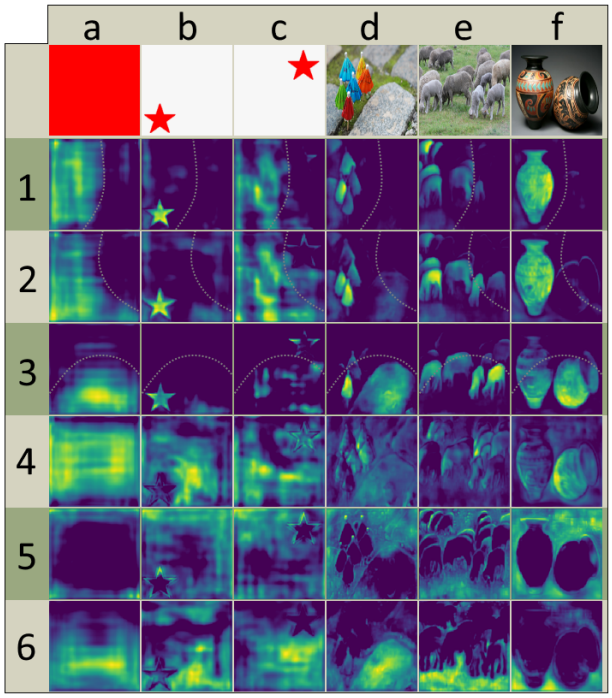


Figure 5: **Prototype Behavior** The activations of the same six prototypes (y axis) across different images (x axis). Prototypes 1-3 respond to objects to one side of a soft, implicit boundary (marked with a dotted line). Prototype 4 activates on the bottom-left of objects (for instance, the bottom left of the umbrellas in image d); prototype 5 activates on the background and on the edges between objects; and prototype 6 segments what the network perceives to be the ground in the image. These last 3 patterns are most clear in images d-f.

multiple prototypes into one, the mask coefficient branch can learn which situations call for which functionality. For instance, in Figure 5, prototype 2 is a partitioning prototype but also fires most strongly on instances in the bottom-left corner. Prototype 3 is similar but for instances on the right. This explains why in practice, the model does not degrade in performance even with as low as $k = 32$ prototypes (see Table 2b). On the other hand, increasing $k$ is ineffective most likely because predicting coefficients is difficult. If the network makes a large error in even one coefficient, due to the nature of linear combinations, the produced mask can vanish or include leakage from other objects. Thus, the network has to play a balancing act to produce the right coefficients, and adding more prototypes makes this harder. In fact, we find that for higher values of $k$, the network simply adds redundant prototypes with small edge-level variations that slightly increase $AP_{95}$, but not much else.

## 4. Backbone Detector

For our backbone detector we prioritize speed as well as feature richness, since predicting these prototypes and coefficients is a difficult task that requires good features to do well. Thus, the design of our backbone detector closely follows RetinaNet [27] with an emphasis on speed.

**YOLACT Detector** We use ResNet-101 [19] with FPN [26] as our default feature backbone and a base image size of $550 \times 550$. We do not preserve aspect ratio in order to get consistent evaluation times per image. Like RetinaNet, we modify FPN by not producing $P_2$ and producing $P_6$ and $P_7$ as successive $3 \times 3$ stride 2 *conv* layers starting from $P_5$ (not $C_5$) and place 3 anchors with aspect ratios $[1, 1/2, 2]$ on each. The anchors of $P_3$ have areas of 24 pixels squared, and every subsequent layer has double the scale of the previous (resulting in the scales $[24, 48, 96, 192, 384]$). For the prediction head attached to each $P_i$, we have one $3 \times 3$ *conv* shared by all three branches, and then each branch gets its own $3 \times 3$ conv in parallel. Compared to RetinaNet, our prediction head design (see Figure 4) is more lightweight and much faster. We apply smooth-$L_1$ loss to train box regressors and encode box regression coordinates in the same way as SSD [30]. To train class prediction, we use softmax cross entropy with $c$ positive labels and 1 background label, selecting training examples using OHEM [39] with a 3:1 neg:pos ratio. Thus, unlike RetinaNet we do not use focal loss, which we found not to be viable in our situation.

With these design choices, we find that this backbone performs better and faster than SSD [30] modified to use ResNet-101 [19], with the same image size.

## 5. Other Improvements

We also discuss other improvements that either increase speed with little effect on performance or increase performance with no speed penalty.

**Fast NMS** After producing bounding box regression coefficients and class confidences for each anchor, like most object detectors we perform NMS to suppress duplicate detections. In many previous works [35, 36, 30, 37, 18, 27], NMS is performed sequentially. That is, for each of the $c$ classes in the dataset, sort the detected boxes descending by confidence, and then for each detection remove all those with lower confidence than it that have an IoU overlap greater than some threshold. While this sequential approach is fast enough at speeds of around 5 fps, it becomes a large barrier for obtaining 30 fps (for instance, a 10 ms improvement at 5 fps results in a 0.26 fps boost, while a 10 ms improvement at 30 fps results in a 12.9 fps boost).

To fix the sequential nature of traditional NMS, we introduce Fast NMS, a version of NMS where every instance can be decided to be kept or discarded in parallel. To do this, we simply allow already-removed detections to suppress other detections, which is not possible in traditional NMS. This relaxation allows us to implement Fast NMS entirely in standard GPU-accelerated matrix operations.

Figure 6: **YOLACT** evaluation results on COCO's `test-dev` set. This base model achieves 29.8 mAP at 33.0 fps. All images have the confidence threshold set to 0.3.

To perform Fast NMS, we first compute a $c \times n \times n$ pairwise IoU matrix $X$ for the top $n$ detections sorted descending by score for each of $c$ classes. Batched sorting on the GPU is readily available and computing IoU can be easily vectorized. Then, we remove detections if there are any higher-scoring detections with a corresponding IoU greater than some threshold $t$. We efficiently implement this by first setting the lower triangle and diagonal of $X$ to 0: $X_{kij} = 0,\ \forall k, j, i \geq j$, which can be performed in one batched `triu` call, and then taking the column-wise max:

$$K_{kj} = \max_{i}(X_{kij}) \qquad \forall k, j \tag{2}$$

to compute a matrix $K$ of maximum IoU values for each detection. Finally, thresholding this matrix with $t$ ($K < t$) will indicate which detections to keep for each class.

Because of the relaxation, Fast NMS has the effect of removing slightly too many boxes. However, the performance hit caused by this is negligible compared to the stark increase in speed (see Table 2a). In our code base, Fast NMS is 11.8 ms faster than a Cython implementation of traditional NMS while only reducing performance by 0.1 mAP. In the Mask R-CNN benchmark suite [18], Fast NMS is 15.0 ms faster than their CUDA implementation of traditional NMS with a performance loss of only 0.3 mAP.

**Semantic Segmentation Loss** While Fast NMS trades a small amount of performance for speed, there are ways to increase performance with no speed penalty. One of those ways is to apply extra losses to the model during training using modules not executed at test time. This effectively increases feature richness while at no speed penalty.

Thus, we apply a semantic segmentation loss on our feature space using layers that are only evaluated during training. Note that because we construct the ground truth for this loss from instance annotations, this does not strictly capture semantic segmentation (i.e., we do not enforce the standard one class per pixel). To create predictions during training, we simply attach a 1x1 *conv* layer with $c$ output channels directly to the largest feature map ($P_3$) in our backbone. Since each pixel can be assigned to more than one class, we use sigmoid and $c$ channels instead of softmax and $c + 1$. This loss is given a weight of 1 and results in a $+0.4$ mAP boost.

## 6. Results

We report instance segmentation results on MS COCO [28] and Pascal 2012 SBD [16] using the standard metrics. For MS COCO, we train on `train2017` and evaluate on `val2017` and `test-dev`.

**Implementation Details** We train all models with batch size 8 *on one GPU* using ImageNet [10] pretrained weights. We find that this is a sufficient batch size to use batch norm, so we leave the pretrained batch norm unfrozen but do not add any extra *bn* layers. We train with SGD for 800k iterations starting at an initial learning rate of $10^{-3}$ and divide by

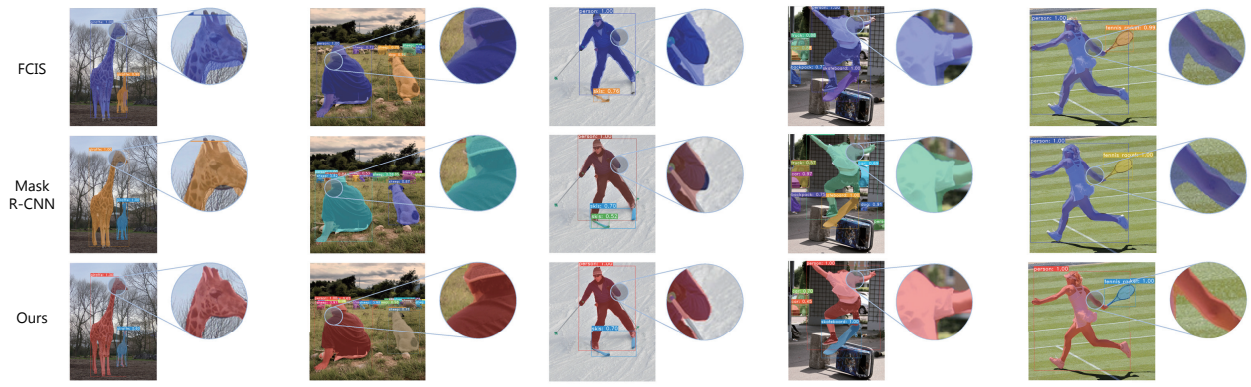


Figure 7: **Mask Quality** Our masks are typically higher quality than those of Mask R-CNN [18] and FCIS [24] because of the larger mask size and lack of feature repooling.

10 at iterations 280k, 600k, 700k, and 750k, using a weight decay of $5\times10^{-4}$, a momentum of 0.9, and all data augmentations used in SSD [30]. For Pascal, we train for 120k iterations and divide the learning rate at 60k and 100k. We also multiply the anchor scales by $4/3$, as objects tend to be larger. Training takes 4-6 days (depending on config) on one Titan Xp for COCO and less than 1 day on Pascal.

**Mask Results** We first compare YOLACT to state-of-the art methods on MS COCO's `test-dev` set in Table 1. Because our main goal is speed, we compare against other single model results with no test-time augmentations. We report all speeds computed on a single Titan Xp, so some listed speeds may be faster than in the original paper.

YOLACT-550 offers competitive instance segmentation performance while at 3.8x the speed of the previous fastest instance segmentation method on COCO. We also note an interesting difference in where the performance of our method lies compared to others. Supporting our qualitative findings in Figure 7, the gap between YOLACT-550 and Mask R-CNN at the 50% overlap threshold is 9.5 AP, while it's 6.6 at the 75% IoU threshold. This is different from the performance of FCIS, for instance, compared to Mask R-CNN where the gap is consistent (AP values of 7.5 and 7.6 respectively). Furthermore, at the highest (95%) IoU threshold, we outperform Mask R-CNN with 1.6 vs. 1.3 AP.

We also report numbers for alternate model configurations in Table 1. In addition to our base $550 \times 550$ image size model, we train $400 \times 400$ (YOLACT-400) and $700 \times 700$ (YOLACT-700) models, adjusting the anchor scales accordingly ($s_x = s_{550}/550 * x$). Lowering the image size results in a large decrease in performance, demonstrating that instance segmentation naturally demands larger images. Then, raising the image size decreases speed significantly but also increases performance, as expected.

In addition to our base backbone of ResNet-101 [19], we also test ResNet-50 and DarkNet-53 [36] to obtain even faster results. If higher speeds are preferable we suggest using ResNet-50 or DarkNet-53 instead of lowering the image size, as these configurations perform much better than YOLACT-400, while only being slightly slower.

Finally, we also train and evaluate our ResNet-50 model on Pascal 2012 SBD in Table 3. YOLACT clearly outperforms popular approaches that report SBD performance, while also being significantly faster.

**Mask Quality** Because we produce a final mask of size $138 \times 138$, and because we create masks directly from the original features (with no repooling to transform and potentially misalign the features), our masks for large objects are noticeably higher quality than those of Mask R-CNN [18] and FCIS [24]. For instance, in Figure 7, YOLACT produces a mask that cleanly follows the boundary of the arm, whereas both FCIS and Mask R-CNN have more noise. Moreover, despite being 5.9 mAP worse overall, at the 95% IoU threshold, our base model achieves 1.6 AP while Mask R-CNN obtains 1.3. This indicates that repooling does result in a quantifiable decrease in mask quality.

**Temporal Stability** Although we only train using static images and do not apply any temporal smoothing, we find that our model produces more temporally stable masks on videos than Mask R-CNN, whose masks jitter across frames even when objects are stationary. We believe our masks are more stable in part because they are higher quality (thus there is less room for error between frames), but mostly because our model is one-stage. Masks produced in two-stage methods are highly dependent on their region proposals in the first stage. In contrast for our method, even if the model predicts different boxes across frames, the prototypes are not affected, yielding much more temporally stable masks.

## 7. Discussion

Despite our masks being higher quality and having nice properties like temporal stability, we fall behind state-of-the-art instance segmentation methods in overall performance, albeit while being much faster. Most errors are caused by mistakes in the detector: misclassification, box

| Method | Backbone | FPS | Time | AP | $AP_{50}$ | $AP_{75}$ | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|---|---|---|
| PA-Net [29] | R-50-FPN | 4.7 | 212.8 | 36.6 | 58.0 | 39.3 | 16.3 | 38.1 | 53.1 |
| RetinaMask [14] | R-101-FPN | 6.0 | 166.7 | 34.7 | 55.4 | 36.9 | 14.3 | 36.7 | 50.5 |
| FCIS [24] | R-101-C5 | 6.6 | 151.5 | 29.5 | 51.5 | 30.2 | 8.0 | 31.0 | 49.7 |
| Mask R-CNN [18] | R-101-FPN | 8.6 | 116.3 | 35.7 | 58.0 | 37.8 | 15.5 | 38.1 | 52.4 |
| MS R-CNN [20] | R-101-FPN | 8.6 | 116.3 | **38.3** | 58.8 | 41.5 | 17.8 | 40.4 | 54.4 |
| YOLACT-550 | R-101-FPN | **33.5** | **29.8** | 29.8 | 48.5 | 31.2 | 9.9 | 31.3 | 47.7 |
| YOLACT-400 | R-101-FPN | 45.3 | 22.1 | 24.9 | 42.0 | 25.4 | 5.0 | 25.3 | 45.0 |
| YOLACT-550 | R-50-FPN | 45.0 | 22.2 | 28.2 | 46.6 | 29.2 | 9.2 | 29.3 | 44.8 |
| YOLACT-550 | D-53-FPN | 40.7 | 24.6 | 28.7 | 46.8 | 30.0 | 9.5 | 29.6 | 45.5 |
| YOLACT-700 | R-101-FPN | 23.4 | 42.7 | 31.2 | 50.6 | 32.8 | 12.1 | 33.3 | 47.1 |

Table 1: **MS COCO [28] Results** We compare to state-of-the-art methods for mask mAP and speed on COCO `test-dev` and include several ablations of our base model, varying backbone network and image size. We denote the backbone architecture with `network-depth-features`, where `R` and `D` refer to ResNet [19] and DarkNet [36], respectively. Our base model, YOLACT-550 with ResNet-101, is 3.9x faster than the previous fastest approach with competitive mask mAP.

| Method | NMS | AP | FPS | Time |
|---|---|---|---|---|
| YOLACT | Standard | **30.0** | 24.0 | 41.6 |
| | Fast | 29.9 | **33.5** | **29.8** |
| Mask R-CNN | Standard | **36.1** | 8.6 | 116.0 |
| | Fast | 35.8 | **9.9** | **101.0** |

(a) **Fast NMS** Fast NMS performs only slightly worse than standard NMS, while being around 12 ms faster. We also observe a similar trade-off implementing Fast NMS in Mask R-CNN.

| $k$ | AP | FPS | Time |
|---|---|---|---|
| 8 | 26.8 | **33.0** | **30.4** |
| 16 | 27.1 | 32.8 | 30.5 |
| *32 | 27.7 | 32.4 | 30.9 |
| 64 | **27.8** | 31.7 | 31.5 |
| 128 | 27.6 | 31.5 | 31.8 |
| 256 | 27.7 | 29.8 | 33.6 |

(b) **Prototypes** Choices for $k$. We choose 32 for its mix of performance and speed.

| Method | AP | FPS | Time |
|---|---|---|---|
| FCIS w/o Mask Voting | 27.8 | 9.5 | 105.3 |
| Mask R-CNN (550 × 550) | **32.2** | 13.5 | 73.9 |
| *fc*-mask | 20.7 | 25.7 | 38.9 |
| YOLACT-550 (Ours) | 29.9 | **33.0** | **30.3** |

(c) **Accelerated Baselines** We compare to other baseline methods by tuning their speed-accuracy trade-offs. *fc*-mask is our model but with 16 × 16 masks produced from an *fc* layer.

Table 2: **Ablations** All models evaluated on COCO `val2017` using our servers. Models in Table 2b were trained for 400k iterations instead of 800k. Time in milliseconds reported for convenience.

| Method | Backbone | FPS | Time | $mAP^r_{50}$ | $mAP^r_{70}$ |
|---|---|---|---|---|---|
| MNC [7] | VGG-16 | 2.8 | 360 | 63.5 | 41.5 |
| FCIS [24] | R-101-C5 | 9.6 | 104 | 65.7 | 52.1 |
| YOLACT-550 | R-50-FPN | **47.6** | **21.0** | **72.3** | **56.2** |

Table 3: **Pascal 2012 SBD [16] Results** Timing for FCIS redone on a Titan Xp for fairness. Since Pascal has fewer and easier detections than COCO, YOLACT does much better than previous methods. Note that COCO and Pascal FPS are not comparable because Pascal has fewer classes.

misalignment, etc. However, we have identified two typical errors caused by YOLACT's mask generation algorithm.

**Localization Failure** If there are too many objects in one spot in a scene, the network can fail to localize each object in its own prototype. In these cases, the network will output something closer to a foreground mask than an instance segmentation for some objects in the group; e.g., in the first image in Figure 6 (row 1 column 1), the blue truck under the red airplane is not properly localized.

**Leakage** Our network leverages the fact that masks are cropped after assembly, and makes no attempt to suppress noise outside of the cropped region. This works fine when the bounding box is accurate, but when it is not, that noise can creep into the instance mask, creating some "leakage" from outside the cropped region. This can also happen when two instances are far away from each other, because the network has learned that it doesn't need to localize far away instances—the cropping will take care of it. However, if the predicted bounding box is too big, the mask will include some of the far away instance's mask as well. For instance, Figure 6 (row 2 column 4) exhibits this leakage because the mask branch deems the three skiers to be far enough away to not have to separate them.

**Understanding the AP Gap** However, localization failure and leakage alone are not enough to explain the almost 6 mAP gap between YOLACT's base model and, say, Mask R-CNN. Indeed, our base model on COCO has just a 2.5 mAP difference between its test-dev mask and box mAP (29.8 mask, 32.3 box), meaning our base model would only gain a few points of mAP even with perfect masks. Moreover, Mask R-CNN has this same mAP difference (35.7 mask, 38.2 box), which suggests that the gap between the two methods lies in the relatively poor performance of our detector and not in our approach to generating masks.

**Acknowledgements** This work was supported in part by ARO YIP W911NF17-1-0410, NSF CAREER IIS-1751206, AWS ML Research Award, Google Cloud Platform research credits, and XSEDE IRI180001.

## References


[1] Shivani Agarwal and Dan Roth. Learning a sparse representation for object detection. In *ECCV*, 2002.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, 2015.

[3] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *CVPR*, 2017.

[4] 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. In *CVPR*, 2018.

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[6] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-sensitive fully convolutional networks. In *ECCV*, 2016.

[7] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016.

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NeurIPS*, 2016.

[9] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[11] Nikita Dvornik, Konstantin Shmelkov, Julien Mairal, and Cordelia Schmid. Blitznet: A real-time deep network for scene understanding. In *ICCV*, 2017.

[12] Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[13] Alireza Fathi, Zbigniew Wojna, Vivek Rathod, Peng Wang, Hyun Oh Song, Sergio Guadarrama, and Kevin Murphy. Semantic instance segmentation via deep metric learning. *arXiv preprint arXiv:1703.10277*, 2017.

[14] Cheng-Yang Fu, Mykhailo Shvets, and Alexander C Berg. Retinamask: Learning to predict masks improves state-of-the-art single-shot detection for free. *arXiv preprint arXiv:1901.03353*, 2019.

[15] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012.

[16] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *ICCV*, 2011.

[17] Adam Harley, Konstantinos Derpanis, and Iasonas Kokkinos. Segmentation-aware convolutional networks using local attention masks. In *ICCV*, 2017.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[20] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *CVPR*, 2019.

[21] Saumya Jetley, Michael Sapienza, Stuart Golodetz, and Philip Torr. Straight to shapes: real-time detection of encoded shapes. In *CVPR*, 2017.

[22] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. Instancecut: from edges to instances with multicut. In *CVPR*, 2017.

[23] Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 2001.

[24] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *CVPR*, 2017.

[25] Xiaodan Liang, Liang Lin, Yunchao Wei, Xiaohui Shen, Jianchao Yang, and Shuicheng Yan. Proposal-free network for instance-level object segmentation. *TPAMI*, 2018.

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *CVPR*, 2017.

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[29] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018.

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[32] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*, 2017.

[33] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *CoRR*, 2016.

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[35] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *CVPR*, 2017.

[36] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv:1804.02767*, 2018.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015.

[38] Xiaofeng Ren and Deva Ramanan. Histograms of sparse codes for object detection. In *CVPR*, 2013.

[39] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016.

[40] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[41] Michael Treml, José Arjona-Medina, Thomas Unterthiner, Rupesh Durgesh, Felix Friedmann, Peter Schuberth, Andreas Mayr, Martin Heusel, Markus Hofmarcher, Michael Widrich, et al. Speeding up semantic segmentation for autonomous driving. In *NeurIPS Workshops*, 2016.

[42] Jonas Uhrig, Eike Rehder, Björn Fröhlich, Uwe Franke, and Thomas Brox. Box2pix: Single-shot instance segmentation by assigning pixels to object boxes. In *IEEE Intelligent Vehicles Symposium*, 2018.

[43] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[44] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super-resolution via sparse representation. *IEEE Transactions on Image Processing*, 2010.

[45] Xiaodong Yu, Li Yi, Cornelia Fermüller, and David Doermann. Object detection using shape codebook. In *BMVC*, 2007.

[46] Tianzhu Zhang, Bernard Ghanem, Si Liu, Changsheng Xu, and Narendra Ahuja. Low-rank sparse coding for image classification. In *ICCV*, 2013.

[47] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnet for real-time semantic segmentation on high-resolution images. In *ECCV*, 2018.