Learning Saliency Propagation for Semi-Supervised Instance Segmentation

Yanzhao Zhou[†1,2], Xin Wang[2], Jianbin Jiao[1], Trevor Darrell[2] and Fisher Yu[2]

[1]University of Chinese Academy of Sciences
[2]UC Berkeley
zhouyanzhao215@mails.ucas.ac.cn, {xinw, trevor}@eecs.berkeley.edu, jiaojb@ucas.ac.cn, i@yf.io

Abstract

Instance segmentation is a challenging task for both modeling and annotation. Due to the high annotation cost, modeling becomes more difficult because of the limited amount of supervision. We aim to improve the accuracy of the existing instance segmentation models by utilizing a large amount of detection supervision. We propose ShapeProp, which learns to activate the salient regions within the object detection and propagate the areas to the whole instance through an iterative learnable message passing module. ShapeProp can benefit from more bounding box supervision to locate the instances more accurately and utilize the feature activations from the larger number of instances to achieve more accurate segmentation. We extensively evaluate ShapeProp on three datasets (MS COCO, PASCAL VOC, and BDD100k) with different supervision setups based on both two-stage (Mask R-CNN) and single-stage (RetinaMask) models. The results show our method establishes new states of the art for semi-supervised instance segmentation. [1]

1. Introduction

Instance segmentation methods [13, 3, 25] have enjoyed great success recently thanks to deep convolutional networks and availability of new datasets [5, 32]. Those methods can be applied in a broad range of applications such as key point detection and 3D cuboid fitting. However, collecting these fine annotations requires prohibitively expensive human effort, preventing the existing frameworks from learning on larger data. This difficulty limits our further study in the instance segmentation problem.

One possible solution is to use the abundant cheaper bounding box annotations to relieve the shortage of segmentation supervision. Some works [15, 20] have explored how to generalize to unseen categories with the weak supervision. However, the segmentation modules usually only get pixel information from only fine segmentation supervision. It is still not well understood how to exploit the new pixel-level information from bounding boxes, even though more box-level supervision can improve the object localization,

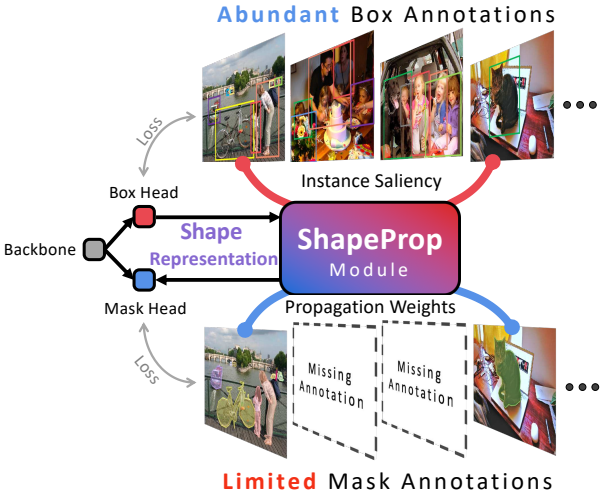

Figure 1: ShapeProp jointly learns from the abundant yet coarse-grained boxes and the fine-detailed yet limited amount of masks to extract a shape representation, which provide strong shape prior to mask head to improve segmentation accuracy and generalization.

In this work, we aim to combine the weekly supervised segmentation information in bounding boxes and full instance segmentation supervision for semi-supervised learning of instance segmentation. The task is a generalization of the "partially supervised instance segmentation" in previous literature [16, 20] which assumes the costly mask annotations are available for a subset of categories. We denote this as category-wise semi-supervision and also consider a realistic image-wise semi-supervision where only a subset of images has masks. The category-wise and image-wise semi-supervision focuses on inter- and intra-class generalization. Note that the considerably cheaper bounding box annotations are available for all object instances.

[†]Work was done while Yanzhao was a visiting scholar at UC Berkeley.
[1]Our source code is available at github.com/ucbdrive/ShapeProp

Semi-supervised instance segmentation is challenging to existing instance segmentation frameworks. Current single-stage (*e.g.*, RetinaMask [10]) or two-stage (*e.g.*, Mask R-CNN family [13, 25]) instance segmentation frameworks do not take full advantage of the existing supervision. They typically use a detection head and a segmentation head to learn from the box and mask supervision separately. So the segmentation head does not explicitly benefit from the abundant box annotations as each box is a coarse-grained representation which does not specify object shape. Also, the segmentation head requires high-level semantics to predict pixel-wise labels, making it difficult to adequately capture data distribution when only limited masks are available.

Our goal is to learn to locate and segment objects from both box and segmentation annotations. We propose ShapeProp, a lightweight network module that can extend existing instance segmentation frameworks to utilize box-level instance supervision for the semi-supervised instance segmentation. ShapeProp exploits the shape prior information hidden inside box and mask annotations to improve accuracy and generalization of existing instance segmentation frameworks, as illustrated in Fig. 1. ShapeProp can learn instance-specific saliency from the abundant box supervision and model the instance-specific latent relationships between pixels from the limited segmentation annotations.

ShapeProp extracts salient regions of the instance from the box detection outputs. Then it propagates salient regions into an intermediate shape representation, referred as *Shape Activation*, which specifies fine-detailed instance shape extents, as shown in Fig. 2. Shape activation indicates the potential object shapes and provides shape prior. We then fuse it with the region features before making final instance segmentation predictions.

Our approach does not use preprocessing steps such as grouping masks in ShapeMask [20], and can be easily integrated and jointly trained with existing instance segmentation frameworks. In contrast to methods [15] based on transfer learning, which only aims to improve inter-class generalization, our approach improves both inter- and intra-class generalization. Furthermore, our method is lightweight and do not introduce heavy computational overhead to the model inference speed.

Extensive experiments show that ShapeProp module brings consistent and significant improvements to baselines, achieving top performance on various benchmarks. For instance, on semi-supervision setting, Mask R-CNN augmented by ShapeProp improves the baseline by 10.8 AP and outperform the state-of-the-art by 2.2 AP. Those results show strong evidence that ShapeProp effectively improves model's segmentation quality and generalization ability.

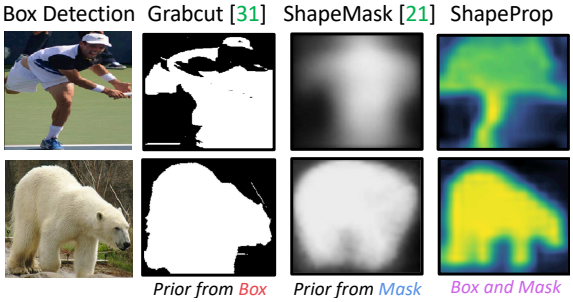

Figure 2: Visualization of extracted shape representation. Previous methods use either box or mask supervision to extract shape prior, while our approach jointly learns from both boxes and masks to extract shape representation.

2. Related Work

Instance segmentation. Instance segmentation is a fundamental task in the computer vision can be roughly categorized into *detection-based* or *grouping-based* approaches. The detection-based methods [11, 6, 12, 21, 4, 13] dominates the state-of-the-art performance in commonly used benchmarks, *e.g.*, COCO [22]. They typically follow a multi-task learning paradigm where a backbone network first extracts spatial features and generates a set of candidate regions, either with region proposal networks [30] or dense anchor boxes [10]. Then a detection head and a segmentation head which composes several convolution -relu layers predicted the accurate box and the segmentation mask inside the region cropped by the detected box. The grouping-based approaches [19, 2, 7, 26, 24, 1, 18] view the instance segmentation as a bottom-up grouping problem. Although great progress has been achieved in the instance segmentation task, most of these works require strong supervision in the form of hand-annotated instance masks for all objects, which limits their application on large-scale datasets.

Weakly supervised instance segmentation. Methods learning with weaker labels try to break this limitation by learning with a weaker form of supervision. [17] leverages the idea that given a bounding box for the target object, one can obtain a pseudo mask label from a grouping-based segmentation algorithm like GrabCut [31]. Pham *et al.* [28] study open-set instance segmentation by using a boundary detector followed by grouping. Zhou *et al.* [33, 34] tackle weakly supervised instance segmentation by exploiting the class peak response of classification networks. Although effective, these approaches do not take advantage of *existing* instance mask labels to achieve better performance.

Learning with limited masks. As opposed to the weakly-supervised setting [17, 33], some recent approaches tackle the partially supervised setting where only box labels (not mask labels) are available for a subset of categories at training time. The model is required to perform instance segmentation on these categories at test time, which requires strong generalization ability. Mask^X R-CNN [15]

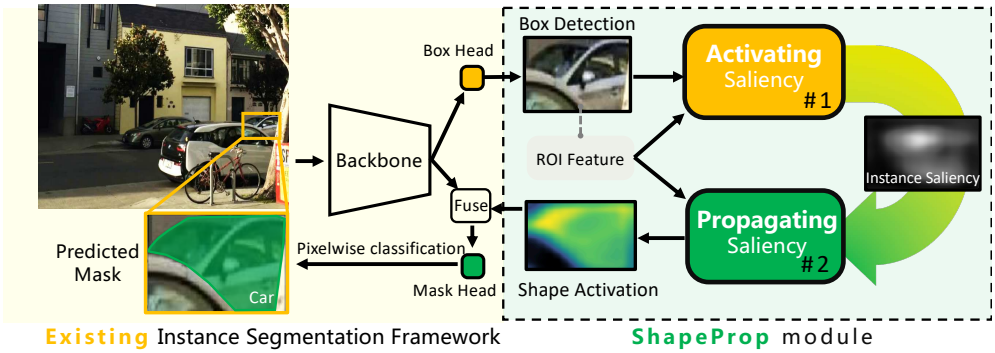

Figure 3: Overall Framework. ShapeProp activates a saliency map for each box detection and propagate the saliency conditioned on the appearance of the object to generate a well-generalized shape representation, *i.e.* Shape Activation, which provide strong shape prior to improve segmentation accuracy and generalization.

tackles the problem by learning to predict weights of mask segmentation branch from the box detection branch. This transfer learning approach shows significant improvement over class-agnostic training; however, performance gap to fully supervised methods remains significant.

Moreover, it only addresses inter-class generalization problems while ignoring intra-class generalization (*i.e.* novel instances from seen categories). ShapeMask [20] is based on a strong assumption that shape bases from existing mask annotations could serve as canonical shapes and generalize to unseen categories. ShapeMask uses the limited mask annotations to extract prior knowledge to help segmentation. Despite its effectiveness, this approach drew class agnostic shape prior from the limited mask annotations and neglect the abundant box annotations, which we argue could provide informative instance-specific saliency.

In this paper, we tackle semi-supervised instance segmentation problem, which includes both category-wise semi-supervision setting (*i.e.*, partially supervised setting), and image-wise semi-supervision setting, *i.e.*, only a small subset of images have masks. Those two settings focus on inter- and intra-class generalization, respectively. In other words, our model generalizes to both objects from unseen categories and novel objects from seen categories.

3. Method

In this section, we first revisit the detection based instance segmentation frameworks as we aim to improve their accuracy and generalization. We then introduce the proposed ShapeProp approach, starting with the process of statistically learning instance-specific saliency via multiple instance learning and followed by the design of propagating saliency to a well-generalized shape representation, referred to as Shape Activation. Finally, we discuss how to integrate Shape Activation to the instance segmentation frameworks. The overall architecture of ShapeProp is illustrated in Fig. 3.

3.1. Learning Saliency Propagation

Activating Saliency. Although a single box annotation does not specify the label for each pixel, they entail information on what an object looks like and provide one weak label for all pixels within the bounding box. We can still learn pixel-level label statistics from the weak supervision after looking at a large number of examples. This is also studied in the context of Multiple Instance Learning (MIL). MIL [27] is a form of weakly supervised learning where training instances are arranged in sets, called bags, and a label is provided for the entire bag instead of an individual instance. Note that in our context, we consider each box as a bag of pixels.

We first construct positive and negative bags from the box detections $\mathcal{B} = \{b_1, b_2, ..., b_N\}$ outputted by the instance segmentation framework where $b_i = (\hat{x}, \hat{y}, \hat{w}, \hat{h}, \hat{c})$ is the i-th detection and N is the number of detected objects in the image. $(\hat{x}, \hat{y}, \hat{w}, \hat{h})$ is the predicted coordinates of the bounding box and $0 \leq \hat{c} < K$ is the predicted class label from K predefined object categories, *e.g.*, car, dog, and person. Let Y_i be the label of a bag $X_i = \{x_1, x_2, ..., x_{\hat{w}*\hat{h}}\}$, where a bag is defined as a set of pixels within the box detection b_i. Each instance (*i.e.*, pixel) x_j corresponds to a binary label y_j which indicates whether the pixel is in the mask of object detected by b_i. We follow the standard MIL assumption that all negative bags contain only negative instances, and positive bags contain at least one positive instance:

$$Y_i = \begin{cases} +1 \; if \; \exists \, y_j \; : \; y_j = +1, \\ -1 \; if \; \forall \, y_j \; : \; y_j = -1. \end{cases} \tag{1}$$

For each category c, we use the box annotations G^c to partition $\mathcal{B}^c$ into positive bags $\mathcal{P}^c = \{p_1, p_2, ..., p_u\}$ and negative bags $\mathcal{N}^c = \{n_1, n_2, ..., n_v\}$ based on the Intersection over Union (IoU) and a threshold t (*e.g.*, $t = 0.5$):

$$\begin{aligned} \mathcal{P}^c &= \{b_i \in \mathcal{B} \; if \; IoU(b_i, g) \; > \; t, \exists \, g \in G^c\}, \\ \mathcal{N}^c &= \{b_i \in \mathcal{B} \; if \; IoU(b_i, g) \; \leq \; t, \forall \, g \in G^c\}. \end{aligned} \tag{2}$$

We do this because positive sample must contain part of the object (*i.e.*, positive bag) as the IoU is high. The negative sample is misaligned with the object (low IoU) thus all pixels are consider invalid (*i.e.*, negative bag). Note that this is different from previous works that utilize MIL to discover class-specific responses in the image for semantic segmentation [29]. We extract object-specific responses for instance segmentation.

We build a weak learner $\mathcal{F}$, which is a lightweight module containing several conv-relu layers. For each sample p_i, $\mathcal{F}$ predicts a map $M \in \mathcal{R}^{h \times w}$ based on the corresponding region feature. In contrast to using pixel-level ground truth, we learn $\mathcal{F}$ using bag-level labels from boxes:

$$L = \sum_{c} (\sum_{p_i \in \mathcal{P}^c} -\sigma(F(\omega(p_i), \theta)) + \sum_{n_j \in \mathcal{N}^c} +\sigma(F(\omega(n_j), \theta))), \tag{3}$$

where L is the loss, θ is the learnable parameters in $\mathcal{F}$, σ is a 2D aggregation operator, *e.g.*, Avg2D, and ω is a region feature pooling operator, *e.g.*, ROIAlign. The learning of $\mathcal{F}$ accumulates several bags of instances to collect pixel-level information statistically, and the predicted M highlights salient regions in the image specific to each detected object and provides rich shape prior information to the subsequent mask head, as shown in Fig. 4.

Propagating Saliency. Object-specific region responses M obtained with MIL highlight only salient regions, we further design a propagation step to make use of the existing limited mask annotations and to recover full object extent from the incomplete object region responses. The motivation is to exploit the relation between pixels to estimate object shape from the salient regions obtained from boxes. Instead of considering it as a pixelwise classification problem, we learn how to propagate messages between deep pixels in the latent feature space. Learning to predict edges between nodes (*i.e.*, deep pixels) for propagation instead of labels of nodes is more effective for the cases with only limited mask annotations. The reason lies in that pixelwise classification depends on high-level semantics and thus requires sufficient supervision to capture the diverse data distribution. Meanwhile, the relationship between pixels, *i.e.*, whether two pixels belong to the same object, can be more reliably inferred with low-level semantics, *e.g.*, similar color, and smooth texture. Intuitively speaking, a model does not need to recognize the semantic category of a banana but still can segment its extent by grouping pixels with a similar color.

The saliency propagation is implemented as a latent space message passing process. As illustrated in Fig. 5, we first use conv blocks (*i.e.*, conv-relu layers) to encode the instance saliency map $M \in \mathcal{R}^{H \times W}$ to latent features $\tilde{M} \in \mathcal{R}^{C \times H \times W}$, where C is the channel dimension (*e.g.*, 16). The encoding extracts features from M while making the subsequent message passing more robust to noise.

For each channel $\tilde{M}_i$, we consider deep pixels as nodes

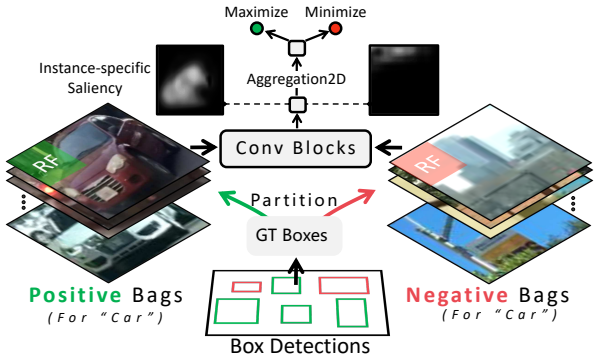

Figure 4: Learning instance-specific saliency from abundant yet coarse-graind box annotations. We partition box detections as bags of receptive fields (RFs) with bag-level labels (*i.e.*, contain instance or not), and statistically learn a weak learner for predicting probability for each RF, resulting a sailency map for each object.

and learn to propagate the message between spatially adjacent nodes. Based on the appearance feature of b, we use conv blocks to predict propagation weights between nodes $W \in \mathbb{R}^{C \times (r \times r) \times (H \times W)}$, where r is a window size (*e.g.*, 3). We then normalize and shuffle the learned propagation weights W to construct location-specific kernels $K_{i,u,v} \in \mathcal{R}^{r \times r}$, where i is the channel dimension and (u, v) is the spatial location and the $\sum_{(p,q)} K^{p,q}_{i,u,v} = 1$. The propagation is an iterative process. At each step, we use K to update the latent features:

$$\tilde{M}_i^{u,v}(t+1) = \sum_{(p,q) \in \mathcal{N}} \tilde{M}_i^{p,q}(t) \cdot K_{i,u,v}^{p-u+\frac{r}{2}, q-v+\frac{r}{2}}, \tag{4}$$

where $\tilde{M}_i^{(u,v)}$ is the feature value at the (u, v) location of i-th channel, and $\mathcal{N}$ is the set of neighbor locations of (u, v). We iterate $max(H, W)$ steps to guarantee that messages can spread over all locations, and each node could absorb information from all other nodes. This iterative process does not introduce significant computation overhead as the spatial size of the region feature is typically very small (*e.g.*, 14), and our efficient GPU implementation computes message passing for all detections simultaneously.

Finally, we use the convolution layer to decode the updated latent features back to a single channel map, referred to as Shape Activation, which combines shape information from both box and mask annotations and specify the extent of the object. Shape Activation serves as an intermediate shape representation that provides strong shape prior to subsequent mask prediction (*i.e.*, mask head). During training, we use binary cross-entropy to compute the reconstruction loss against the ground truth mask. For the semi-supervised setting, we only calculate the loss in the small subset of images that have mask annotations. Note that learning of saliency propagation is in a class-agnostic manner that allows it to accumulate common knowledge from all existing

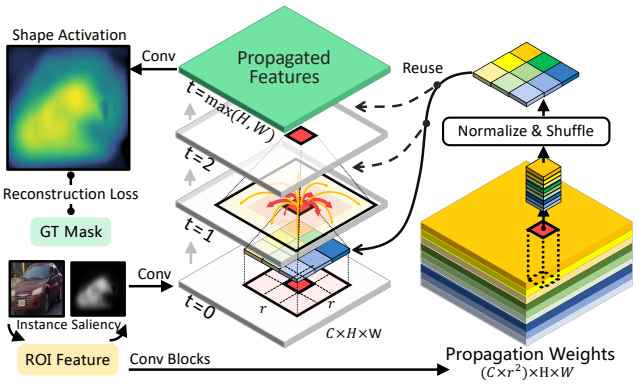

Figure 5: Saliency Propagation. We learn to predict latent relationship between pixels conditioned on instance apperance and utilize iterative message passing to propagate saliency into a intermediate shape representation, referred to as Shape Activation.

masks and effectively generalize to novel categories.

3.2. Intergrating Shape Representation

We integrate the learned Shape Activation $\mathcal{S}$ into existing instance segmentation frameworks by concatenating it with the input region features before feeding into the mask head. The additional input channels provide strong prior information of objects' possible shapes; thus can significantly simplify the task of learning mask prediction, *i.e.*, pixel-level classification, and allows mask head to focus on capturing fine-detailed information. Experimentally, we show that the Shape Activation guided segmentation not only significantly improves generalization ability (10.8% AP improvements on COCO's partially supervised setting), but also benefit segmentation quality even when mask annotations are sufficient (1.4% AP_{75} improvements on COCO's fully supervised setting).

4. Experiments

We evaluate the proposed ShapeProp method on popular instance segmentation benchmarks including COCO [22], PASCAL-VOC [9] and BDD100K [32]. We report standard metrics, that is, mask AP, AP50, AP75, and AP, for small/medium/large objects, following the evaluation protocol in previous works [13, 16, 20].

In Sec. 4.1, we test our method on the category-wise semi-supervision setting (*i.e.*, "partially supervised" setting in previous literature [16, 20]). The significant improvements over baselines (10.8% AP) and the new state-of-the-art results indicate ShapeProp's capability to learn from limited masks and generalize to unseen categories. In Sec. 4.2, we benchmark on BDD100K, where only a subset of images have mask annotations. We augment both the single-stage and two-stages instance segmentation frameworks [10, 13] with ShapeProp and show consistent improvements over baselines. This indicates the effectiveness of ShapeProp's intra-class generalization. We further show that our approach can improve strong baselines that are trained using full mask annotations from the dataset. In Sec. 4.3, statistical analyses show that ShapeProp can learn high-quality shape representation. Finally, we perform ablation studies to investigate our model design further.

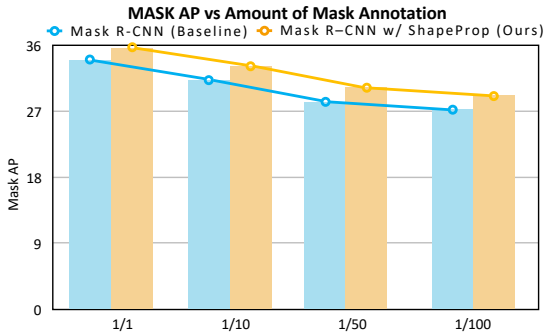

Figure 6: Generalization with less data. ShapeProp generalizes well down to 1/100 of the training data and it consistently improves the baseline.

4.1. Generalization to unseen categories

Experiment setup. The experiments are set up following [15, 20]. We split the COCO17's 80 categories into VOC (20) vs. Non-VOC (60). The VOC categories are also present in PASCAL VOC [8]. At training time, models have access to the bounding boxes of all classes, but the masks only come from either VOC or Non-VOC categories. The performance upper bounds are set by the oracle models that have access to masks from all categories. In this section, our training set is COCO train2017, and the comparison with other methods is done on val2017 Non-VOC/VOC.

We build our models by plugging the ShapeProp module into the representative Mask R-CNN framework [13]. In order to evaluate across categories, we use the class-agnostic setting, which considers all object classes as the foreground class. We implement the model with two backbones, *i.e.*, ResNet50-FPN and ResNet101-FPN [14, 23]. We use the same training parameters as the baseline.

Numerical results. It can be seen in table 1, Mask R-CNN equipped with ShapeProp improves the baseline by a significant margin (*e.g.*, 34.4% vs 23.9% for non-voc $\rightarrow$ voc and 30.4% vs 19.2% for voc $\rightarrow$ non-voc). Our model with ResNet50-FPN backbone outperforms the state-of-the-art ShapeMask [20] that is build on a stronger backbone (ResNet101-FPN. Our ShapeProp module improves segmentation by fully exploiting shape knowledge from the box and mask annotations; thus, it can also benefit from other advances in deep learning *i.e.*, stronger backbone. Switching to ResNet101-FPN backbone improves yields a top result on these benchmarks (*e.g.*, 2.2% AP higher than state of the art ShapeMask [20]).

Backbone	Method	non-voc → voc: test on $B = \{\text{voc}\}$						voc → non-voc: test on $B = \{\text{non-voc}\}$					
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
R-50-FPN	Mask R-CNN trained w/ GrabCut [17]	19.5	46.2	14.2	4.7	15.9	32.0	19.5	39.2	17.0	6.5	20.9	34.3
	Mask^X R-CNN [15]	28.9	52.2	28.6	12.1	29.0	40.6	23.7	43.1	23.5	12.4	27.6	32.9
	Mask R-CNN (Baseline) [13]	23.9	42.9	23.5	11.6	24.3	33.7	19.2	36.4	18.4	11.5	23.3	24.4
	Mask R-CNN w/ ShapeProp (Ours)	**34.4**	**59.6**	**35.2**	**13.5**	**32.9**	**48.6**	**30.4**	**51.2**	**31.8**	**14.3**	**34.2**	**44.7**
	fully supervised (Oracle) [13]	37.5	63.1	38.9	15.1	36.0	53.1	33.0	53.7	35.0	15.1	37.0	49.9
R-101-FPN	Mask R-CNN trained w/ GrabCut [17]	19.6	46.1	14.3	5.1	16.0	32.4	19.7	39.7	17.0	6.4	21.2	35.8
	Mask^X R-CNN [16]	29.5	52.4	29.7	13.4	30.2	41.0	23.8	42.9	23.5	12.7	28.1	33.5
	ShapeMask [20]	33.3	56.9	34.3	17.1	38.1	45.4	30.2	49.3	31.5	16.1	38.2	28.4
	Mask R-CNN (Baseline) [13]	24.7	43.5	24.9	11.4	25.7	35.1	18.5	34.8	18.1	11.3	23.4	21.7
	Mask R-CNN w/ ShapeProp (Ours)	**35.5**	**60.5**	**36.7**	15.6	33.8	**50.3**	**31.9**	**52.1**	**33.7**	14.2	35.9	**46.5**
	fully supervised (Oracle) [13]	38.5	64.4	40.4	18.9	39.4	51.4	34.3	54.7	36.3	18.6	39.1	47.9

Table 1: Performance on COCO2017's category-wise semi-supervision setting which focuses on inter-class generalization. At the top, non-voc→voc means "train on masks in non-voc, test on masks in voc", and vice versa. Equipping ShapeProp module to class-agnostic Mask R-CNN gives 10.8 AP improvements and beat ShapeMask by 2.2 AP, showing strong evidence that ShapeProp can significantly improve existing models' accuracy and generalization.

Generalization with less data. To further validate the generalization ability of ShapeProp with less training data, we train the ResNet50-FPN based models on full categories using only 1/10, 1/50, 1/100 of the data. As can be seen in Fig. 6, our approach generalizes well down to 1/100 of the training data, and it consistently outperforms the baseline (Mask R-CNN without ShapeProp).

Qualitative results. Fig. 7 gives qualitative examples from the non-voc → voc setting. It can be seen in the second row, the baseline method failed to segment the novel category "bicycle" as no masks for this category is available during training. However, adding ShapeProp to the baseline significantly improves the segmentation quality. It can also be seen in the bottom row that Mask R-CNN predicts a broken mask for the "cow" instance. In contrast, Mask R-CNN with ShapeProp model segment it correctly.

4.2. Generalization to novel instances

Experiment setup. We further benchmark upon the BDD100K dataset [32], which is the largest and most diverse driving video dataset. Due to the extensive human efforts required for labeling detailed instance segmentation, only a subset of BDD100K provides mask annotations. The dataset fits naturally with our semi-supervised setting. We fuse the annotations of BDD100K's instance segmentation and detection to build a data contains 67k images with box annotations, among which 7k images have mask annotations. We test models on BDD100K's val. set (1k images).

We build our models by plugging the proposed ShapeProp module into two representative detection based instance segmentation frameworks, *i.e.*, Mask R-CNN (two-stages method), and RetinaMask (single-stage method). We compare with the joint learning version of Mask R-CNN. It learns from all images and only compute the loss for segmentation head when mask annotation is available. We also compare with Grabcut Mask R-CNN [17] and Progressive Mask R-CNN, which use Grabcut post-processing and the pretrained annotator model to obtain pseudo masks from box annotations. All models are based on the ResNet-50 FPN backbone and are trained via standard SGD optimization with LR 0.01 and batch size 12.

Numerical results. As shown in Tab. 2, the mask APs for both single-stage and two-stage baselines are significantly improved when more box annotations are available, *i.e.*, Mask RCNN (24.5 vs 21.6) and RetinaMask (24.4 vs 20.0). However, the model trained with Grabcut pseudo masks performs even worse than the baseline, indicating the shape representation quality from Grabcut is not good enough. It can be seen in Tab. 2, equipping the proposed ShapeProp module bridges the learning of box and mask and further improves the model's segmentation ability (26.2% vs 24.5%). This shows ShapeProp can effectively exploit shape prior hidden inside the annotations to enhance the quality of segmentation. Moreover, it can be seen in Tab. 3, equipping ShapeProp also improves the strong baselines trained with fully supervised setting where all masks are available during training. This indicates that fully exploiting annotations can improve segmentation quality even when masks are sufficient.

Inference time. The proposed ShapeProp module is a lightweight module built on top of the convolution blocks. The message passing operation for propagating saliency is efficiently implemented as matrix dot production. Therefore, overall ShapeProp module does not introduce heavy computation overhead (0.35 vs 0.39 s/img on 2080 Ti).

Qualitative results. In Fig. 8, we visualize examples of box detection, instance saliency, shape activation, and mask predictions from models with or without ShapeProp. It can be seen from the left sample, multiple cars are occluded in the detected region, and the baseline model failed to segment the correct one. In contrast, ShapeProp find the object-specific salient parts and further predict propagation

Figure 7: Visualization of Mask R-CNN (w/ or wo/ ShapeProp)'s results on novel categories. Results demonstrate that plugging ShapeProp module leads to significant better segmentation quality.

Training data (# annotations)	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Comment
BDD100k-Instseg (7k w/ masks, 7k w/ boxes)	RetinaMask [10]	20.0	37.6	18.1	7.6	27.0	42.3	Single-stage baseline
	Mask R-CNN [13]	21.6	40.5	20.5	9.3	28.8	45.4	Two-stage baseline
BDD100k (7k w/ masks, 67k w/ boxes)	RetinaMask [10]	24.4	44.7	22.5	9.8	32.6	51.5	Joint training
	RetinaMask w/ ShapeProp	**26.1**	**46.8**	**24.9**	**10.7**	**34.7**	**56.3**	ShapeProp augmented (Ours)
	Mask R-CNN [13]	24.5	45.4	21.6	10.1	33.1	48.3	Joint training
	Grabcut Mask R-CNN [17]	21.0	41.0	19.5	8.3	27.0	40.7	Grabcut limited mask
	Progressive Mask R-CNN	24.8	45.4	23.0	10.0	33.0	52.7	Progressive learning
	Mask R-CNN w/ ShapeProp	**26.2**	**48.4**	**23.5**	**11.4**	**34.2**	**53.0**	ShapeProp augmented (Ours)

Table 2: Performance on BDD100K's image-wise semi-supervision setting that focuses on intra-class generalization. ShapeProp improves the accuracy and generalization of single-stage (RetinaMask) and two-stage (Mask R-CNN) framework.

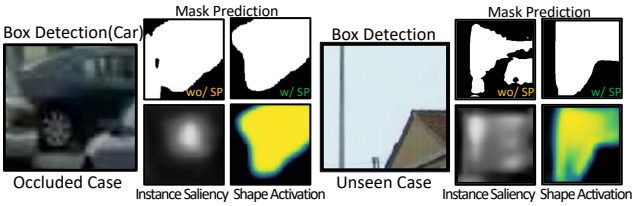

Figure 8: Visualization of samples in BDD100K datasets.

weights conditional on the visual appearance to propagate salient regions into a high-quality shape activation. Based on the shape activation's strong shape prior, the mask prediction of our model is significantly improved. The right sample gave a case when detection failed, *i.e.*, a background region is mis-detected as a car. Despite this background region is unseen from the training data, our model gives a more reasonable shape representation and mask prediction while the baseline outputs a mask with broken pieces. Note the model is designed to estimate the shape of the centered salient "object" in the box. This further demonstrates the strong generalization of our approach.

4.3. Analysis and ablation studies

Statistical Analysis. We analyze the quality of the extracted shape prior with respect to object size, demonstrating that our approach can effectively extract shape prior from the box and mask annotations. Shape activation is assigned to GT masks and judged by measuring the best overlap metric. To be considered a perfect

Dataset	Method	AP	AP_{50}	AP_{75}
VOC12	Mask R-CNN [13]	29.7	54.1	29.6
(1k images)	**ShapeProp (Ours)**	**30.6**	**53.2**	**32.0**
BDD100K	Mask R-CNN [13]	21.6	40.5	20.5
(7k images)	**ShapeProp (Ours)**	**22.7**	**42.2**	**21.8**
COCO17	Mask R-CNN [13]	34.2	56.4	36.7
(120k images)	**ShapeProp (Ours)**	**35.7**	**56.9**	**38.2**

Table 3: Comparison under fully supervised setting. ShapeProp-augmented models consistently improve AP.

Setting	AP	AP_{50}	AP_{75}
baseline	21.6	40.5	20.5
baseline + More head params	21.4	40.3	19.5
baseline + ShapeProp (Channel-agnostic)	22.4	42.0	20.5
baseline + ShapeProp	**22.7**	**42.2**	**21.8**
baseline + More boxes	24.5	45.4	21.6
baseline + More boxes + ShapeProp	**26.2**	**48.4**	**23.5**

Table 4: Ablation studies of model design on BDD100K.

shape activation that completely coincide with a GT mask, the IoU between the predicted shape activation M and GT masks $\mathcal{T}$ must be close to 100% as computed using the metric $\max_{\theta, T_i \in \mathcal{T}} \frac{area(f_b(M,\theta) \cap T_i)}{area(f_b(M,\theta) \cup T_i)}$, where the function $f_b(M, \theta) = M \geq \theta$ produces the best matching binary instance masks based on the probabilistic shape activation over a set of threshold values $\theta \in (0, 1)$. Note that this metric is robust to the absolute value range of prediction and is suitable for evaluating probabilistic activation maps.

We visualize the density of the best overlap for instance-specific saliency and the shape activation obtained by message passing to see whether the activation can cover object instances of different sizes. Fig. 9 (left) shows that saliency samples clustered in the area where the best overlap value is around 60% and failed to cover large objects. In contrast, in Fig. 9 (right), most of the data points have relatively high best overlap IoUs and perform very well on both small and large objects. It can be seen that the message-passing design in our ShapeProp significantly improves the quality of the extracted shape prior. The underlying reason is that multiple instance learning typically captures the salient regions of the instance; however, it failed to recover the full extent. The message passing utilizes the limited number of ground truth masks to learn how to refine the saliency map and recover such extent in a class agnostic and well-generalized manner, thus considerably improve the quality.

Ablation Study. We perform ablation studies on the BDD100K dataset to validate some specific designs of our method, Tab. 4. The baseline model is a Mask R-CNN with a ResNet50-FPN backbone. The first to sixth rows correspond to models trained on the instseg split of BDD100K, which contains 7k images for training, and all instances in the training data have both box and mask annotations. The last two rows are trained with the overall BDD100K data,

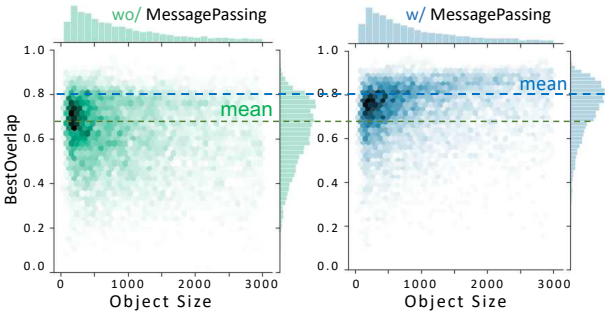

Figure 9: Shape activation accuracy. Message passing process propagates saliency to better cover object extent.

including its official detection and instseg split, which has 67k images in total among it; only 7k images have masks. The setting of the second row increases the number of channels used in Mask R-CNN's mask head to make the number of learnable parameters be the same as the ShapeProp-augmented counterpart. The AP of this setting is even slightly worse than the baseline, which validates that the gain of ShapeProp doesn't come from the increasing of model parameters. Comparing the fifth to the third rows, the better AP indicates that predicting propagation weights for each channel of the latent space instead of sharing the same weights for all channels can lead to performance improvements. The sixth row, which uses additional bounding box annotations from the BDD100K detection split has 2.9% AP improvements (24.5% vs 21.6%). Meanwhile, the last row, which is augmented by the ShapeProp module, has 3.5% AP gains (26.2 % vs 22.7%) over the baseline and 4.6% AP improvements over the plain baseline of the first row. This clearly shows that the ShapeProp module can better utilize the hidden shape prior to the additional box annotations to benefit the segmentation quality.

5. Conclusions

We developed a lightweight network module, ShapeProp, which can be plugged into existing instance segmentation frameworks to tackle the challenging semi-supervised instance segmentation task. ShapeProp extracts a well-generalized shape representation from the joint learning of the abundant yet coarse-grind box supervision and the fine-detailed yet limited amount of mask annotations. Such shape representation hypothesis possible object shape and specify detailed instance boundaries that provide strong shape prior to the subsequent mask prediction, which allows the learning of strong instance segmentation model based on limited mask annotations. We extensively test ShapeProp on popular benchmarks, including COCO, PASCAL VOC, and BDD100K. The results indicate that ShapeProp-augmented frameworks consistently outperform baseline by a significant margin, establishing states-of-the-art for semi-supervised instance segmentation.

References

[1] Anurag Arnab and Philip HS Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *CVPR*, volume 1, page 5, 2017. 2

[2] Min Bai and Raquel Urtasun. Deep watershed transform for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5221–5229, 2017. 2

[3] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2018. 2

[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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[6] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3150–3158, 2016. 2

[7] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551*, 2017. 2

[8] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009. 5

[9] Mark Everingham, S. M. Ali Eslami, Luc J. Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136, 2015. 5

[10] 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. 2, 5, 7

[11] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014. 2

[12] Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Boundary-aware instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5696–5704, 2017. 2

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 1, 2, 5, 6, 7, 8

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[15] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 5, 6

[16] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to Segment Every Thing. In *CVPR*, 2018. 1, 5, 6

[17] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1665–1674, 2017. 2, 6, 7

[18] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. Instancecut: from edges to instances with multicut. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5008–5017, 2017. 2

[19] Shu Kong and Charless C Fowlkes. Recurrent pixel embedding for instance grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9018–9028, 2018. 2

[20] 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. 1, 2, 3, 5, 6

[21] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4438–4446, 2017. 2

[22] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 2, 5

[23] Tsung-Yi Lin, Piotr Dollár, Ross B Girshick, Kaiming He, Bharath Hariharan, and Serge J Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017. 5

[24] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. Sgn: Sequential grouping networks for instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3496–3504, 2017. 2

[25] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8759–8768, 2018. 1, 2

[26] Yiding Liu, Siyu Yang, Bin Li, Wengang Zhou, Jizheng Xu, Houqiang Li, and Yan Lu. Affinity derivation and graph merge for instance segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 686–703, 2018. 2

[27] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. In *Advances in neural information processing systems*, pages 570–576, 1998. 3

[28] Trung Pham, Vijay BG Kumar, Thanh-Toan Do, Gustavo Carneiro, and Ian Reid. Bayesian semantic instance segmentation in open set world. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–18, 2018. 2

[29] Pedro O Pinheiro and Ronan Collobert. Weakly supervised semantic segmentation with convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, page 6, 2015. 4

[30] 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 (NIPS)*, pages 91–99, 2015. 2

[31] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, pages 309–314. ACM, 2004. 2

[32] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 5, 6

[33] Yanzhao Zhou, Yi Zhu, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Weakly supervised instance segmentation using class peak response. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[34] Yi Zhu, Yanzhao Zhou, Huijuan Xu, Qixiang Ye, David Doermann, and Jianbin Jiao. Learning instance activation maps for weakly supervised instance segmentation. In *CVPR*, 2019. 2