This WACV 2021 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

The Devil is in the Boundary: Exploiting Boundary Representation for Basis-based Instance Segmentation

Myungchul Kim
KAIST
gritycda@kaist.ac.kr

Sanghyun Woo
KAIST
shwoo93@kaist.ac.kr

Dahun Kim
KAIST
mcahny@kaist.ac.kr

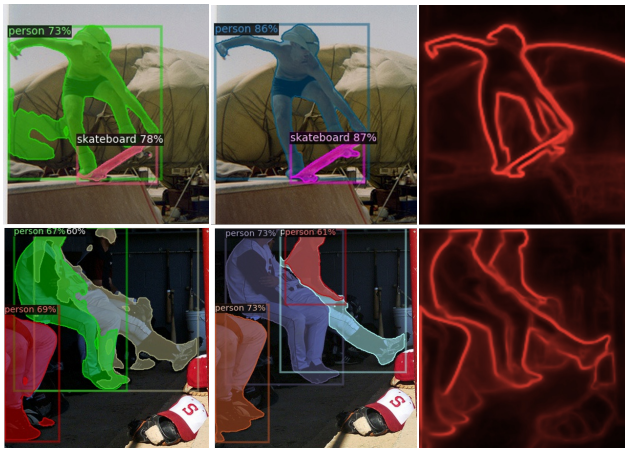
In So Kweon
KAIST
iskweon@kaist.ac.kr

Abstract

*Pursuing a more coherent scene understanding towards real-time vision applications, single-stage instance segmentation has recently gained popularity, achieving a simpler and more efficient design than its two-stage counterparts. Besides, its global mask representation often leads to superior accuracy to the two-stage Mask R-CNN which has been dominant thus far. Despite the promising advances in single-stage methods, finer delineation of instance boundaries still remains unexcavated. Indeed, boundary information provides a strong shape representation that can operate in synergy with the fully-convolutional mask features of the single-stage segmenter. In this work, we propose Boundary Basis based Instance Segmentation(**B2Inst**) to learn a global boundary representation that can complement existing global-mask-based methods that are often lacking high-frequency details. Besides, we devise a unified quality measure of both mask and boundary and introduce a network block that learns to score the per-instance predictions of itself. When applied to the strongest baselines in single-stage instance segmentation, our B2Inst leads to consistent improvements and accurately parse out the instance boundaries in a scene. Regardless of being single-stage or two-stage frameworks, we outperform the existing state-of-the-art methods on the COCO dataset with the same ResNet-50 and ResNet-101 backbones.*

1. Introduction

Deep learning has been widely used in various vision tasks [18, 14, 34, 45, 32, 44, 21, 43, 2]. Most deep learning based instance segmentation methods follow the *detect-then-segment* paradigm. It first detects bounding boxes and then segments the instance mask in each bounding box. However, this dominant paradigm has two fundamental issues. First, it is step-wise, heavily relying on the accurate bounding box proposals. Second, it produces low-resolution and coarse masks, resulting from only exploiting the region-specific information, *i.e.*, RoI-pooled features. To remedy these issues, recent studies attempt to incorporate a global representation called 'basis' into the framework. They consider combining the global image-level information, *i.e.*, basis, with the original region-specific information in a single-stage and fully convolutional manner.



(a) BlendMask (b) Our B2Inst (c) Our boundary basis

Figure 1. **BlendMask vs B2Inst**: (a) Coarse boundaries appeared in the instance segmentation results of BlendMask [6]. (b) Our proposed boundary-aware masks are more precise in delineating complex shapes like human poses (top) and overlapping instances (bottom). (c) Our holistic image boundary representation helps resolving ambiguities in a challenging scene.

Despite their great effectiveness and flexibility, we argue that *boundary* is a key representation for segmentation that is still missing in this basis framework. Indeed, image boundaries are an important cue for recognition; Humans can recognize objects from sketches alone as instances in natural images are bounded by the edges. In the existing single-stage methods, the global mask representations are not explicitly learned in a supervised way, but loosely trained by a signal from the last mask output. However, such implicit learning of mask representations often ignores object shape and boundary information. This can lead to

coarse and indistinct segmentation results because the pixels near boundaries are hard to be classified, as shown in Fig. 1. We are not the first one to exploit boundaries for instance segmentation. Recently, Boundary-preserving Mask RCNN [11] proposed a boundary prediction head alongside the mask head of the two-stage Mask RCNN. They have shown that the boundary representation can provide a more precise object location and mitigates its coarse and low-resolution predictions. Motivated by this recent observation, we extend the idea in the context of the recent single-stage instance segmentation. First, while the RoI-wise head of Boundary-preserving [11] learns the boundary information for a single object at a time, we propose to learn holistic image-level instance boundaries. Intuitively, heavy occlusions and shape complexities are better revealed at a whole image extent. We thus argue that learning the global boundary has its distinct advantages in segmenting individual objects in a scene, and will have a great synergy with the global mask representations in modern single-stage segmenters. We name our proposed method as Boundary Basis based Instance Segmentation(**B2Inst**), and this new boundary basis is explicitly trained with the boundary ground-truths, which can be automatically generated from the given mask annotations. All masks and boundary representations are then cropped and composited together to form the final instance masks. Second, instead of using the instance boundary information for representation learning only during the training, we use the model to learn a novel instance-wise boundary-aware mask score, which can be mainly utilized in test-time to prioritize high-quality masks. It is a generalized version of [20] and considers the agreement between the predicted masks and boundaries as well. It is a unified measure of both mask and boundary quality.

To sum up, we explicitly use image and instance boundaries as an additional important learning cue. We find that both the proposed holistic image boundary representation and the boundary-aware mask score operate complementarily to each other, providing better localization performance and help to resolve ambiguities between overlapping objects, as shown in Fig. 1. We apply our *B2Inst* to the strongest single-stage instance segmentation methods, *i.e.*, BlendMask [6] and YOLACT [4], and achieve consistent improvements on the challenging COCO benchmark.

The main contributions of this paper are as follows.

1. This is the first time to explore the boundary representation for accurate basis-based instance segmentation.
2. We propose *B2Inst* which incorporates the boundary representation both on the global and the local view. At an image-level, we propose to learn holistic boundaries to enhance global basis representation. In an instance-level, we design a boundary-aware mask score. It captures both the mask and boundary quality at the same time and can prioritize high-quality mask predictions at test time. We show that both are complementary to each other in building an accurate segmenter.
3. We conduct extensive ablation studies to verify the effectiveness of our proposals. To further show that our algorithm is general, we apply it to the two strong basis-based instance segmentation frameworks and demonstrate consistent improvements both visually and numerically. Finally, by augmenting the BlendMask framework with our method, we achieve new state-of-the-art results.

2. Related Work

2.1. Instance segmentation

Most existing instance segmentation algorithms can be categorized into either two-stage or one-stage approaches. The two-stage methods first detect the bounding box for each object instance and then predict the segmentation mask inside the cropped box region. Among many previous studies [15, 12, 23, 26, 8, 33, 22], Mask RCNN [17] is the most representative of this category, and it employs a proposal generation network and RoIAlign feature pooling strategies to obtain fixed-sized features of each proposal. Further improvements have been made to boost its accuracy; PANet [30] introduces bottom-up path augmentation to enrich FPN features, and Mask scoring RCNN [20] addresses the misalignment between the confidence score and localization accuracy of predicted masks. These two-stage approach has dominated in state-of-the-art performances thus far in instance segmentation. However, these methods require RoI-wise feature pooling and head operations which make their inference quite slow, limiting the use in real-time applications.

Single-stage approaches typically aim at faster inference speed by avoiding proposal generation and feature pooling operations. In particular, bottom-up methods view the instance segmentation as a label-then-cluster problem. Instead of assigning RoI proposals, it produces pixel-wise predictions of cues such as directional vectors [29], pairwise affinity [31], watershed energy [1], and embedding learning, and then group object instances from the cues in the post-processing stage. There are local-area-based methods that output instance masks on each local region directly [9, 39] without dependency on box detection or grouping post-processing. However, the single-stage approaches thus far have been lagging behind in accuracy compared to the two-stage counterparts, especially on the challenging dataset like COCO. The latest global-area-based methods [6, 4, 37] have broken the record, often outperforming the Mask RCNN [17]. They first generate intermediate FCN feature maps, called 'basis', then assemble the extracted basis features to form the final masks for each instance.

In this work, we especially focus on the latest category

of basis-based methods. More detailed review will be provided in Sec. 3.

2.2. Boundary learning for instance segmentation

Boundary detection has been a fundamental computer vision task as it provides an important cue for recognition [7, 16, 42, 10, 40, 35, 19, 3]. Accurate boundary localization is able to explicitly contribute to the mask prediction for segmentation. CASENet [41] presents a challenging task of category-aware boundary detection. InstanceCut [23] adopts boundaries to partition semantic segmentation into instance-level segmentation. However, it involves expensive super-pixel extractions and grouping computations. Zimmermann et al. propose boundary agreement head [46] to focus on instance boundaries with an auxiliary edge loss. Boundary preserving Mask RCNN [11] proposes to predict instance-level boundaries to augment the mask head. These two-stage methods focus on the RoI-wise boundary of a single object, and thus lack a holistic view of image boundaries which can often resolve ambiguities in multiple overlapping objects and complex scenes.

3. Background

Basis-based instance segmentation: There has been a wave of studies on single-stage instance segmentation, enabled by recent advances in one-stage detection such as FCOS, which outperforms the two-stage counterpart in accuracy. These dense instance segmentation methods generate a set of basis mask representations, called *basis*, by a fully convolutional network. In parallel, there is another task of predicting instance-level parameters given the detected boxes, that are used in combining the basis features into the final per-instance masks. Recent basis-based instance segmentation methods follow this general pipeline [4, 5, 37, 6, 36].

First, YOLACT [4] presents to predict 32 global bases and the according instance-specific *scalar* coefficients. The final mask is then computed by a linear combination among the bases. After YOLACT, the community has pushed either in the direction of developing rich instance features or designing effective assembly methods. BlendMask [6], SipMask [5], and CenterMask [37] improved the assembly parameters, by extending from the scalar to *spatial* 2D coefficient matrices. CondInst [36] considers the *dynamic filter weights* as implicit instance-specific representation. The mask is then predicted by a standard sliding window operation over the basis features, with the predicted convolution weight filters.

In this work, we take one step further by investigating the under-explored boundary information for both the basis and instance-specific features. We first propose to learn holistic boundaries to enhance the global basis representation. Then, we design a boundary-aware mask score to unify and jointly learn the mask and boundary quality of an object instance. Putting together, our method consistently improves over the latest basis-based instance segmentation methods, which implies the generality and orthogonality of our methods to the existing ones.

4. Exploiting Boundary Representation

Overall architecture of our *B2Inst* method is shown in Fig. 2. It consists of four parts: (1) backbone for feature extraction, (2) instance-specific detection head, (3) global basis head, and (4) mask scoring head. The final masks are computed by combining the instance-specific information from the detection head and the global image information from the basis head. We note that Fig. 2 shows the BlendMask instantiation.

4.1. Learning holistic image boundary

Standard basis head: The basis head takes FPN [27] features as input and produces a set of K global basis features. It consists of four consecutive 3×3 convolution layers followed by an upsampling and the last layer that reduces the channel-size to K. The previous basis head [4, 5, 6, 36] is supervised by the last mask loss only, and thus the basis representations are obtained in an unsupervised manner.

Missing piece - image boundary: There is a notable recent study in two-stage framework that improves instance segmentation by learning RoI-wise boundary information. Different from them, we propose to learn a holistic boundary of all instances in a scene, instead of a single instance one by one. Overlapping objects and complex shapes are more clearly captured given the whole image context (see Fig. 1). Moreover, the boundary supervision comes at no cost from the given mask annotations.

Boundary ground-truths: We use the Laplacian operator to generate soft boundaries from the binary mask ground-truths. The boundaries are then thresholded at 0 and converted into final binary maps.

Objective function: We use three different loss functions in total: 1) binary cross-entropy loss, 2) dice loss, and 3) boundary loss [13]. Especially, the boundary loss is proposed to predict sharp boundaries at test time without any sophisticated post-processing.

4.2. Boundary-aware mask scoring

Mask scoring R-CNN [20] proposes a mask IoU scoring module instead of using the classification score to evaluate the mask, which can improve the quality of mask results. The mask IoU (Jaccard index) evaluates the agreement between the mask prediction and ground-truth by focusing only on their *area*, without considering their *shapes*. The discrepancy between the mask's IoU score and visual quality can be found in the top-performing segmentation

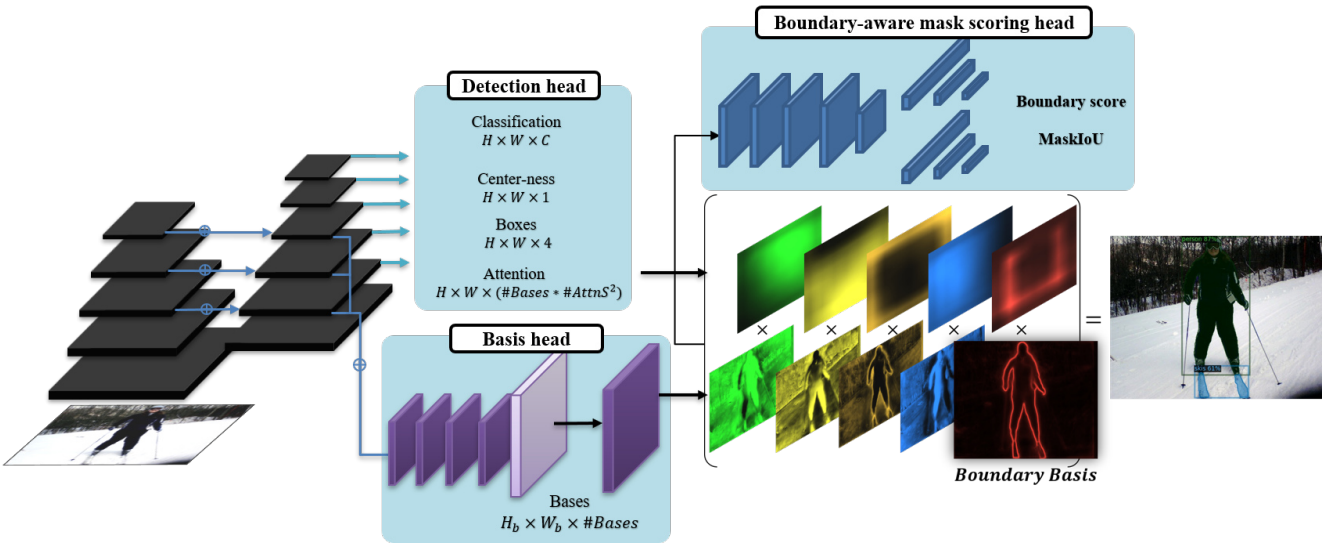


Figure 2. **Overview of our method.** The figure shows the BlendMask [6] instantiation. It generates a set of basis mask representations. In parallel, it predicts instance-level parameters given the detected boxes, *i.e.*, attention maps, which are used in combining the basis features into the final per-instance masks. On top of this strong basis-based instance segmentation framework, we incorporate the boundary representation both on the global and the local view. At an image-level, the model learns to produce a holistic image boundary, improving the global basis features. At an instance-level, the model predicts a boundary-aware mask score, which captures both the mask and boundary quality jointly and can prioritize high-quality mask predictions at test time.

method [6] (Fig. 4-a). Between the mask predictions and ground-truths, their *areas* agree on most pixels, while their *boundaries* do not. Such a problem becomes even more severe when there are overlapping objects and fine details (row 1 and 3), yet is not able to be quantified by the current mask IoU metric. To address this issue, we devise a mask score that can respect both the IoU and boundary quality.

Boundary score: We define $S_{boundary}$ to evaluate the boundary agreement between mask prediction M_{pred} and ground-truth M_{gt}. We use a Dice metric [13] which is insensitive to the number of foreground and background pixels, and thus can have balanced importance on the boundary pixels. The Dice metric also has a range from 0 to 1, which makes it a comparable scoring function to the IoU metric. To compute $S_{boundary}$, we first identify the boundaries in the predicted and ground-truth masks by filtering them with a Laplacian kernel Δf as $B_{pred} = \Delta f(M_{pred})$ and $B_{gt} = \Delta f(M_{gt})$. Then, the Dice metric is computed between the two mask boundaries as in Eq. (1).

$$S_{boundary} = \frac{2\sum_i^{h\times w} B^i_{pred} B^i_{gt} + \epsilon}{\sum_i^{h\times w} (B^i_{pred})^2 + \sum_i^{h\times w} (B^i_{gt})^2 + \epsilon} \quad (1)$$

where i stands for the i-pixel and ϵ is a soft-term to avoid division of zero, and h and w are the height and width of the predicted bounding box size.

S_{mask} can be divided into 1) a classification problem matching the correct class and 2) a problem of regressing the maskIoU in the proposal of the corresponding class. In the first task, S_{class} is the confidence score of which class the bounding box proposal corresponds to. S_{class} is in charge of the classification task at the proposal stage. The other tasks, S_{IoU} is regressed using an additional header in MS-RCNN [20] The object mask and its boundary are complementary, so we can easily change either one to another. Therefore, we designed a simple network that regresses S_{IoU} and S_{Dice} at the same time. This part will be covered in the next paragraph.

Scoring head: We introduce a unified mask scoring head that can predict the mask scores during training. As we want to incorporate the boundary score $S_{boundary}$ with the conventional mask IoU score S_{IoU}, our score prediction head is trained to regress the two scores at a time. To fully exploit the correlation between mask boundary and IoU information, it is designed to be a shared header of 4 convolution layers with last fully-connected branches for each S_{IoU} and $S_{boundary}$ regression tasks. The inputs of our scoring head are a concatenation of predicted mask M_{pred}, boundary B_{pred}, and RoI-pooled FPN features F_{RoI}, all in the shape of $\mathbb{R}^{h\times w}$. We use the ground-truth bounding boxes during the training of our boundary-aware mask scoring head.

Score definition at inference: At inference, we define the mask score S_{mask} using S_{IoU} and $S_{boundary}$ that are predicted from our score head. We follow the standard inference procedure [17]; First, the top-k scoring proposals from the detection head are fed into our scoring head, then the

mask scores are produced. Finally, our mask confidence score is computed by multiplying the S_{IoU} and $S_{boundary}$ with the standard classification score S_{mask} of the bounding box as follows:

$$S_{mask} = S_{class} \cdot \sqrt{S_{IoU} \cdot S_{boundary}}. \quad (2)$$

5. Experiment

In this section, we conduct experiments to analyze our two major proposals in *B2Inst* : 1) holistic boundary basis and 2) boundary-aware mask scoring. Specifically, we perform ablation studies in Sec. 5.1. We then incorporate our proposals into the two strong basis-based instance segmentation frameworks in Sec. 5.2. Finally, we evaluate our mask results quantitatively and qualitatively, compared with the state-of-the-art baselines in Sec. 5.3.

Dataset & Evaluation Metric We present experimental results on the MS COCO instance segmentation benchmark [28]. All the experiments conducted here follow standard COCO 2017 setup, using 115k images as $train$ split, 5k images as $validation$ split. We conduct detailed ablation experiments on the validation split using BlendMask [6] framework. We then report the final COCO mask AP on 20k $test\text{-}dev$ split. We use the official evaluation server as there are no labels released to the public. For the evaluation metric, we use standard COCO-style AP. The results include AP at different scales, AP_S, AP_M, AP_L.

5.1. Ablation study

5.1.1 Impact of two major proposals

The main components of our framework design are the holistic boundary basis and boundary-aware mask-scoring. The ablation results are summarized in Table. 1.

Holistic Boundary basis (HBB) To see the holistic boundary feature's impact, we add it to the original global basis representations. The performance of *B2Inst* with additional boundary information improves by 0.4 AP overall. The improvement of AP_{75} and AP_S are relatively big, implying that the boundary information effectively guides the model to predict more precise masks in challenging cases such as precise localization or small-object segmentation. Note that the cost of obtaining the ground truth boundary is negligible.

Boundary-aware mask scoring (BS + MS) We generalize the original mask score [20] to consider the agreement between the predicted masks and boundaries. In this way, the scoring metric can capture both the mask (*area*) and boundary (*shape*) quality jointly. As shown in the table, both scoring metrics can push the performance, demonstrating their efficacy. Further improvement comes when using both together (by 0.6 AP), meaning that both the mask and boundary quality are crucial when evaluating the mask quality at test time.

HBB	BS	MS	AP	AP50	AP75	APs	APm	APl
			35.8	56.4	38.1	16.9	38.8	51.1
✓			**36.2**	56.5	38.5	17.6	39.3	51.3
✓	✓		**36.3**	55.3	39.4	16.8	39.3	52.3
✓		✓	**36.6**	55.5	39.6	17.3	39.7	51.9
✓	✓	✓	**36.8**	55.7	39.9	17.2	39.9	52.5

Table 1. **Impact of two major proposals.** We verify the efficacy of two major proposals, holistic boundary basis, and boundary-aware mask scoring. HBB, BS, and MS denote Holistic Boundary Basis, Boundary Dice Coefficient Scoring, and Mask IoU Scoring, respectively. We observe that both proposals contribute to the final performance advanced.

In a brief conclusion, we show that the holistic boundary basis improves global basis representation. It especially resolves the ambiguities in challenging cases. Moreover, the proposed boundary-aware mask scoring well prioritizes high-quality masks at test time, demonstrating the efficacy of considering both the area and shape quality jointly. We show that both proposals operate complementarily to each other, contributing to the final performance improved consistently.

Design	AP	AP50	AP75	APs	APm	APl
Baseline	35.8	56.4	38.1	16.9	38.8	51.1
(a)	35.7	56.2	38.0	17.3	38.6	50.9
(b)	36.0	56.4	38.4	16.4	39.0	51.8
(c)	36.0	56.5	38.5	16.5	39.0	51.6
(d)	36.0	56.2	38.4	16.7	38.7	51.5
(e)	**36.2**	56.5	38.5	17.6	39.3	51.3

Table 2. **Basis head design choices.** We compare five different ways of configuring the basis head to predict boundary representation. The details of each configurations are elaborated in Sec. 5.1.2. The experiments are conducted on COCO $val2017$.

5.1.2 Basis head design choices

In this experiment, we compare five different ways of designing the basis head to predict boundary representation (see Fig. 3). The results are shown in Table. 2. We describe each configuration in the following:

- **(a)** The boundary loss is applied locally, *i.e.*, on the RoI pooled basis feature.
- **(b)** Following Boundary Mask-RCNN [11], multi-scale feature fusion is adopted. In addition, the boundary loss is applied globally, *i.e.*, on the whole basis feature.
- **(c)** (b) + The boundary prediction is explicitly used for the final per-instance mask composition.
- **(d)** No multi-scale feature fusion is adopted. The boundary loss is applied globally.

Design	FPS	Mask						BBOX					
		AP	AP50	AP75	APs	APm	APl	AP	AP50	AP75	APs	APm	APl
YOLACT	45.3	28.1	46.2	29.0	9.7	30.4	46.4	30.2	50.6	31.6	14.5	32.3	44.2
B2Inst-YOLACT	42.5	**30.1**	48.8	31.6	10.7	32.2	48.1	**31.1**	53.0	32.3	14.6	34.5	44.8
BlendMask*	10.5	35.8	56.4	38.1	16.9	38.8	51.1	40.2	59.2	43.6	24.2	43.8	51.5
B2Inst-BlendMask*	9.7	**36.7**	55.7	39.9	17.2	39.9	52.5	**40.2**	59.0	43.7	23.5	43.6	52.3

Table 3. **Combining with existing basis-based methods.** In order to show that our approach is general, we apply our proposals to the two different state-of-the-art basis-based methods, YOLACT [4] and BlendMask [6]. The experiments are conducted on COCO $val2017$. We see that our method consistently improves the baseline performances clearly with only a slight difference in *FPS*.

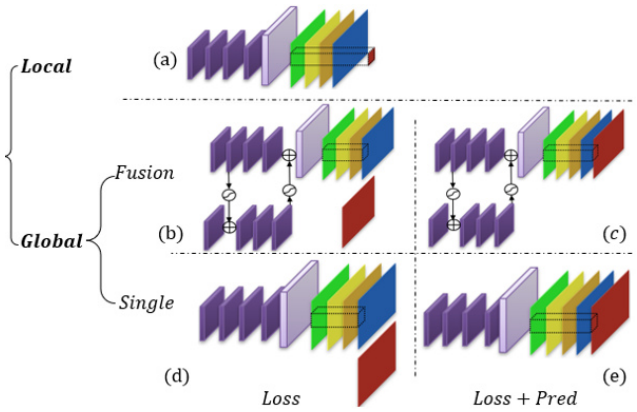


Figure 3. **Five different ways to design the basis head.** *Local* and *Global* indicate whether we use boundary supervision locally or globally. *Fusion* and *Single* indicate whether we use multi-scale feature fusion [11] or not. *Loss + Pred* denotes that we use both the boundary supervision and its prediction. Green, Yellow, Orange, Blue, and Red images represent the original basis. The Red image represents the additional image boundary basis.

- **(e)** (d) + The boundary prediction is explicitly used for the final per-instance mask composition.

We observe that the global boundary supervision is important and also utilizing the boundary prediction as an additional basis representation is effective. We do not see any improvement when using the multi-scale feature aggregation. The configuration (e) produces the best performance, and we use this setup in the following experiments.

5.2. Combining with existing basis-based methods

We see our proposals are general, thus can be easily applied to the existing basis-based methods. In this experiment, we combine our two major proposals with two different basis-based approaches: YOLACT [4] and BlendMask [6]. The results are summarized in Table. 3. We observe that our method consistently improves the performance of all the baselines. More specifically, our proposals improve those frameworks by 2.1 AP and 0.9 AP, respectively. The positive results imply that the boundary representation is indeed a fundamental representation for recognition and is missing in the previous basis-based frameworks. We show that the proposed concept of both the holistic boundary representation and the boundary-aware mask score effectively addresses the issue.

5.3. Main results

5.3.1 Quantitative results

We compare our method with several state-of-the-art instance segmentation methods, including both the two-stage [12, 26, 8, 17, 20, 11] and one-stage [9, 36, 24, 6, 4]. All methods are trained on COCO $train2017$ and evaluated on COCO $test\text{-}dev2017$. Table. 4 shows the results. Without bells and whistles, we can surpass the existing state-of-the-art methods with the same ResNet-50 and ResNet-101 backbones. In particular, Mask R-CNN [17] and MS R-CNN [20] achieve overall mask AP scores of 34.6 and 35.8 in R-50 backbone, respectively. The recently introduced BMask R-CNN [11] obtain mask AP scores of 35.9. Our method achieves 36.8 on the same R-50 backbone.

In case of single-stage methods, YOLACT-700 [4] and TensorMask [9] obtain a mask AP of 31.2 and 37.3 in R-101 backbone, respectively. CondInst [36] and BlendMask* achieve a mask AP score of 39.1 and 39.5, respectively. CenterMask [25] and PointRend [24] obtain the best results among the existing single-stage methods with a mask AP score of 39.8. Under the same setting, *i.e.*, input size and backbone, our method outperforms CenterMask [25] and PointRend [24] with a healthy margin (+ 1.0).

5.3.2 Qualitative results

Improved Boundary Prediction Fig. 4 shows visual mask results of BlendMask [6] and ours on the COCO validation set. The first column (a) is the BlendMask result, and the second column (b) is our result. The third column (c) is the corresponding holistic image boundary bases that are used in (b) to predict final masks. The first image is monochrome photography with the salt and pepper noise. BlendMask is vulnerable to this noise, and thus the object and its surroundings are not delineated accurately. In the second image, BlendMask misses predicting the gull's wing on the boundary area. In the third image, BlendMask fails to predict the kite area correctly due to the overlapping similar textures. On the other hand, we can clearly see that our result has a much better mask prediction quality, especially in the boundary area.

Backbone	Method	epochs	Aug	AP	AP50	AP75	APs	APm	APl
R-50_FPN	**Two-Stage:**								
	Mask R-CNN	12		34.6	56.5	36.6	15.4	36.3	49.7
	MS R-CNN	12		35.8	56.5	38.4	16.2	37.4	51.0
	BMask R-CNN	12		35.9	57.0	38.6	15.8	37.6	52.2
	Single-Stage:								
	TensorMask	72	✓	35.5	57.3	37.4	16.6	37.0	49.1
	CondInst	12	✓	35.9	56.4	37.6	18.4	37.9	46.9
	PointRend	12		36.3	-	-	-	-	-
	Blendmask	12		34.3	55.4	36.6	14.9	36.4	48.9
	Blendmask*	12	✓	35.8	56.4	38.3	16.9	38.8	51.1
	B2Inst-Blend	12	✓	**36.8**	55.7	39.9	17.2	39.9	52.5
	B2Inst-Blend	36	✓	**39.0**	58.7	42.3	19.9	42.0	55.4
R-101_FPN	**Two-Stage:**								
	MNC	12		24.6	44.3	24.8	4.7	25.9	43.6
	FCIS	12		29.2	49.5	-	-	-	-
	FCIS++	12		33.6	54.5	-	-	-	-
	MaskLab	12		35.4	57.4	37.4	16.9	38.3	49.2
	Mask R-CNN	12		36.2	58.6	38.4	16.4	38.4	52.1
	MS R-CNN	12		37.5	58.7	40.2	17.2	39.5	53.0
	BMask R-CNN	12		37.7	59.3	40.6	16.8	39.9	54.6
	Mask R-CNN*	36	✓	38.8	60.9	41.9	21.8	41.4	50.5
	Single-Stage:								
	YOLACT-700	54	✓	31.2	50.6	32.8	12.1	33.3	47.1
	TensorMask	72	✓	37.3	59.5	39.5	17.5	39.3	51.6
	CondInst	36	✓	39.1	60.9	42.0	21.5	41.7	50.9
	Centermask*	36	✓	39.8	-	-	21.7	42.5	52.0
	PointRend	36	✓	39.8	-	-	-	-	-
	Blendmask	36	✓	38.4	60.7	41.3	18.2	41.5	53.3
	Blendmask*	36	✓	39.5	61.1	42.3	19.5	42.9	56.2
	B2Inst-Blend	12	✓	**38.9**	58.4	42.1	18.4	42.3	55.6
	B2Inst-Blend	36	✓	**40.8**	61.0	44.2	19.8	44.1	58.3

Table 4. **Quantitative comparisons with the state-of-the-art methods.** We use MS-COCO $test\text{-}dev2017$ to obtain the final performances. We compare our method with several state-of-the-art instance segmentation methods, including both the two-stage and single-stage. * indicates the reproduced performance in our $Detectron2$ [38] platform. 'Aug.' denotes using multi-scale training with shorter side range in [640, 800].

Improved Mask Prediction Fig. 5 compares the visual mask results of YOLACT [4], BlendMask [6] and ours. The first column (a) is the YOLACT result, and the second column (b) is the result of *B2Inst*-YOLACT. Similarly, the third column (c) is the BlendMask result, and the fourth column (d) is the result of *B2Inst*-BlendMask. Again, we observe that our proposals successfully guide the baseline models to better capture the target object.

6. Conclusion

In this paper, we propose *B2Inst* which learns an image boundary representation to enhance instance segmentation. We build upon the single-stage framework which is drawing much attention nowadays due to its speed and higher accuracy. Our *B2Inst* learns a boundary basis representation by an explicit boundary prediction learning. In particular, the boundary basis is assembled together with the existing mask basis representations to compute the final per-instance masks, providing a better understanding of occlusions and complex shapes in a scene. In addition, we present a quality metric that evaluates both the mask and boundary prediction of each object and trains our model to predict the unified score that can be used in mask ranking at test time. Benefiting from these two proposed components, our model delineates instance shapes more accurately, especially for heavily occluded and complex scenes. When applied to BlendMask and YOLACT, which are the strongest single-stage methods, we exhibit consistent improvements over them both visually and numerically. Our model outperforms the state-of-the-art segmenters both for single-stage and two-stage frameworks. We hope our initial findings of the importance of boundary representation invigorate the follow-up studies.

Acknowledgment

This work was supported by the Technology Innovation Program: Development of core technology for advanced locomotion/manipulation based on high-speed/power robot platform and robot intelligence (No.10070171) funded By the Ministry of Trade, industry & Energy(MI, Korea)



Figure 4. **Qualitative Comparisons on the Boundary Area.** We compare the mask results of (a) BlendMask [6] and (b) Our proposed *B2Inst*. (c) Our learned image boundary representation. We can clearly see that our proposals resolve ambiguities in a challenging scene and produce more accurate masks.

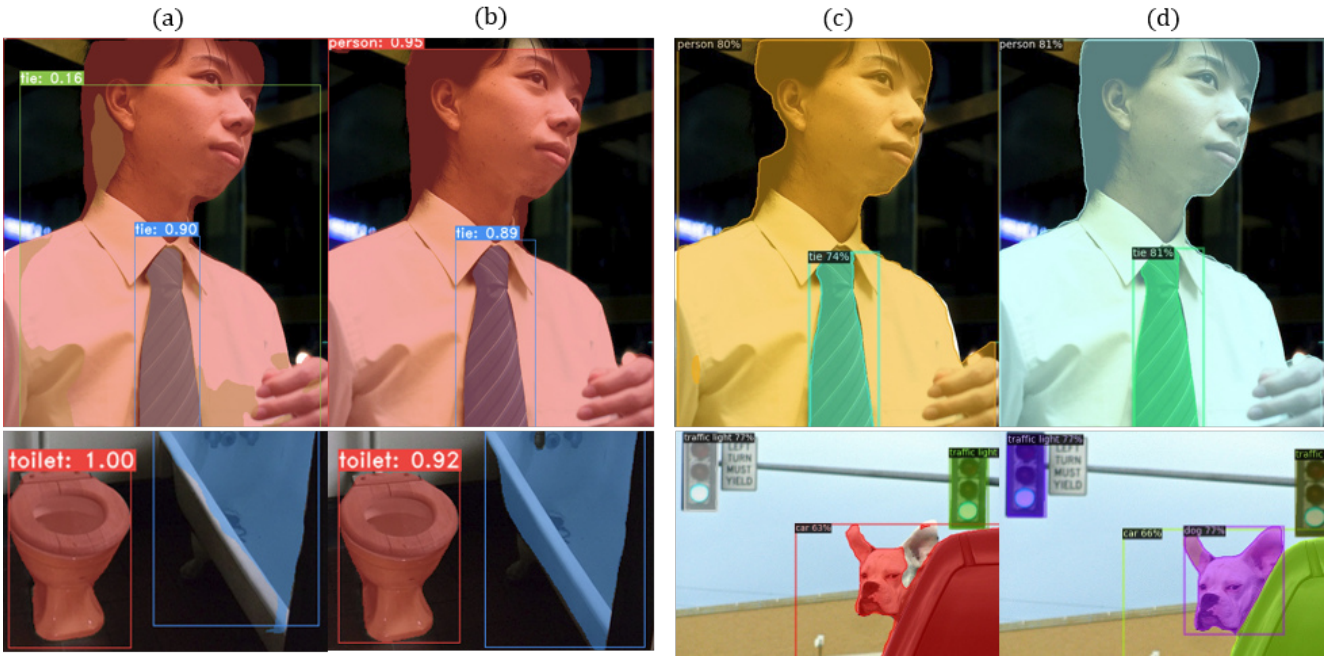


Figure 5. **Qualitative Comparisons between the Baselines [4, 6] and Ours on the COCO validation set.** (a) is the YOLACT result and (b) is the result of *B2Inst*-YOLACT. Similarly, (c) is the BlendMask result, and (d) is the result of *B2Inst*-BlendMask.

References

[1] 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

[2] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Revisiting batch normalization for improving corruption robustness. *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. 1

[3] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Semantic segmentation with boundary neural fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3602–3610, 2016. 3

[4] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, 2019. 2, 3, 6, 7, 8

[5] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast instance segmentation. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2020. 3

[6] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 8573–8581, 2020. 1, 2, 3, 4, 5, 6, 7, 8

[7] Liang-Chieh Chen, Jonathan T Barron, George Papandreou, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4545–4554, 2016. 3

[8] 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 *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 4013–4022, 2018. 2, 6

[9] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, pages 2061–2069, 2019. 2, 6

[10] Feng Cheng, Cheng Chen, Yukang Wang, Heshui Shi, Yukun Cao, Dandan Tu, Changzheng Zhang, and Yongchao Xu. Learning directional feature maps for cardiac mri segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 108–117. Springer, 2020. 3

[11] Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2020. 2, 3, 5, 6

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

[13] Ruoxi Deng, Chunhua Shen, Shengjun Liu, Huibing Wang, and Xinru Liu. Learning to predict crisp boundaries. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 562–578, 2018. 3, 4

[14] Ross Girshick. Fast r-cnn. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[15] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 297–312. Springer, 2014. 2

[16] Zeeshan Hayder, Xuming He, and Mathieu Salzmann. Boundary-aware instance segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 5696–5704, 2017. 3

[17] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *Proc. of Int'l Conf. on Computer Vision (ICCV)*, Oct 2017. 2, 4, 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[19] Qin Huang, Chunyang Xia, Wenchao Zheng, Yuhang Song, Hao Xu, and C-C Jay Kuo. Object boundary guided semantic segmentation. In *Asian Conference on Computer Vision*, pages 197–212. Springer, 2016. 3

[20] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 6409–6418, 2019. 2, 3, 4, 5, 6

[21] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[22] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[23] Alexander Kirillov, Evgeny Levinkov, Bjoern Andres, Bogdan Savchynskyy, and Carsten Rother. Instancecut: from edges to instances with multicut. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 5008–5017, 2017. 2, 3

[24] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 9799–9808, 2020. 6

[25] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[26] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 2359–2367, 2017. 2, 6

[27] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2014. 5

[29] 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

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

[31] 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

[32] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[33] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *Proc. of European Conf. on Computer Vision (ECCV)*, pages 75–91. Springer, 2016. 2

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. of Neural Information Processing Systems (NIPS)*, 2015. 1

[35] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5229–5238, 2019. 3

[36] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Proc. of European Conf. on Computer Vision (ECCV)*, 2020. 3, 6

[37] Yuqing Wang, Zhaoliang Xu, Hao Shen, Baoshan Cheng, and Lirong Yang. Centermask: Single shot instance segmentation with point representation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3

[38] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 7

[39] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 12193–12202, 2020. 2

[40] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1857–1866, 2018. 3

[41] Zhiding Yu, Chen Feng, Ming-Yu Liu, and Srikumar Ramalingam. Casenet: Deep category-aware semantic edge detection. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 5964–5973, 2017. 3

[42] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *European Conference on Computer Vision*, pages 489–506. Springer, 2020. 3

[43] Chaoning Zhang, Philipp Benz, Dawit Mureja Argaw, Seokju Lee, Junsik Kim, Francois Rameau, Jean-Charles Bazin, and In So Kweon. Resnet or densenet? introducing dense shortcuts to resnet. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021. 1

[44] Chaoning Zhang, Francois Rameau, Junsik Kim, Dawit Mureja Argaw, Jean-Charles Bazin, and In So Kweon. Deepptz: Deep self-calibration for ptz cameras. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020. 1

[45] Chaoning Zhang, Francois Rameau, Seokju Lee, Junsik Kim, Philipp Benz, Dawit Mureja Argaw, Jean-Charles Bazin, and In So Kweon. Revisiting residual networks with nonlinear shortcuts. In *British Machine Vision Conference (BMVC)*, 2019. 1

[46] Roland S Zimmermann and Julien N Siems. Faster training of mask r-cnn by focusing on instance boundaries. *Computer Vision and Image Understanding*, 188:102795, 2019. 3