ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240537

Boundary-Enhanced Instance Segmentation

Fangyuan Zhang[a,*], **Tianxiang Pan**[b], **Yu-Wing Tai**[c] **and Bin Wang**[d]

[a, b, d]School of Software, Tsinghua University, China
[a, b, d]Beijing National Research Center for Information Science and Technology (BNRist), China
[c]Department of Computer Science, Dartmouth College

Abstract. Despite significant progress in instance segmentation, recent solutions still fall short of boundary accuracy especially for overlapping instances of the same category. In this paper, we propose a novel *boundary-enhanced instance segmentation (BEIS)* framework that explicitly models the feature relationships across object boundaries for high-quality instance segmentation. Specifically, BEIS generates boundary-enhanced features using both *intra-mask* and *cross-image boundary discrimination learning*. The intra-mask boundary discrimination learning (IBDL) employs pixel-level discrimination learning to disentangle pixel representations along boundaries. The cross-image boundary discrimination learning (CBDL) learns a boundary-aware feature bank from training data to further boost the performance. Thus, CBDL can take advantage of boundary relations across images to enhance the quality of segmented boundaries. To focus on hard-to-segment boundaries, we propose an adaptive sampling strategy to automatically construct discriminative pairs in regions with high possibilities of confusion. Extensive experiments show BEIS outperforms on various datasets.

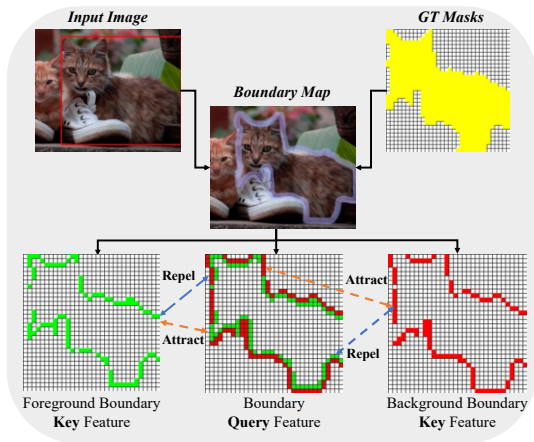

Figure 1. BEIS utilizes boundary discrimination to enhance the distinctiveness of boundary regions, resulting in accurate boundary segmentation.

1 Introduction

Instance segmentation is a pivotal and intricate task in the field of computer vision, finding applications in diverse domains like autonomous driving, scene comprehension, and image manipulation. Prominent approaches for instance segmentation, like Mask R-CNN [9], typically follow a two-stage framework. Initially, a detector generates precise bounding boxes, followed by a parallel segmentation branch that predicts binary masks for instances inside within these boxes. Despite the commendable performance achieved by Mask R-CNN and its derivatives [5, 15, 34] in large-scale instance segmentation tasks, they still encounter challenges in producing highly accurate boundaries, an essential aspect for top-quality instance segmentation. Our observation shows that this limitation is mainly due to confusing boundary features. Current methods struggle with precise segmentation when foreground and background share similar visual or semantic traits, as shown in Figure 2.

This work introduces a novel paradigm for instance segmentation called *Boundary-Enhanced Instance Segmentation (BEIS)*, aimed at enhancing boundary discrimination within fully supervised settings. Illustrated in Figure 1, the BEIS approach leverages boundary information from binary mask features. For each pixel within a mask, positive and negative samples are constructed from neighboring regions within the same mask. Similar to Sup-CL [14], a multi-positive InfoNCE [3] loss is applied, pulling pixel features with matching boundary labels while pushing features with distinct semantics. This process, referred to as intra-mask boundary discrimination learning (IBDL), operates exclusively within a mask. However, for intricate boundary relationships, solely exploring intra-mask connections is insufficient. To address this, we extend the concept to cross-image boundary discrimination learning (CBDL), introducing foreground and background boundary feature banks. Contrastive keys from these banks facilitate inter-image boundary relation exploration. An adaptive query and key selection method is proposed to prioritize challenging boundary pairs with similar semantic and spatial attributes. Both IBDL and CBDL effectively segregate boundary features from confusing neighbors, adeptly managing mask conflicts in overlapping instances without considerable computational cost.

While BEIS is simple yet impactful, its potential is curtailed by coarse mask features lacking intricate boundary details. Thus, to fully harness BEIS capabilities, we propose a refinement framework incorporating multi-stage boundary discrimination. Benefitting from its streamlined architecture, BEIS demands minimal computation resources during inference. Through extensive experiments on diverse datasets including MS-COCO [21], CityScapes [6], and LVIS [8], our approach consistently achieves state-of-the-art performance.

In summary, the key contributions of this paper are:

- Introduction of the BEIS, encompassing intra-mask boundary discrimination learning (IBDL) and cross-image boundary discrimination learning (CBDL), effectively learns boundary-aware dis-

* Email: zhangfy19@mails.tsinghua.edu.cn

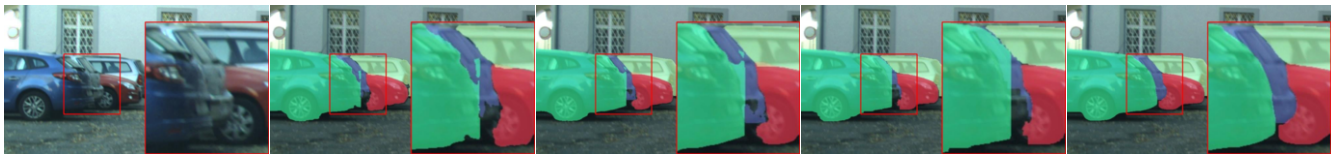

Figure 2. Instance Segmentation on Cityscapes val-set by PointRend [15], RefineMask [34], PatchDCT [30], and BEIS. BEIS produces significantly accurate results at unrecognizable boundary regions.

criminative features for high-quality instance segmentation.

- Introduction of the adaptive query and key selection strategies that emphasize confusing boundary regions.
- Incorporation of a multi-scale implementation within BEIS, enabling robust boundary segmentation across different scales for improved accuracy.
- Extensive experimentation across multiple datasets, showcasing the BEIS's groundbreaking performance in instance segmentation. Notably, BEIS outperforms Mask R-CNN on the COCO and Cityscapes datasets, with AP improvements of 3.3 and 5.4, along with boundary AP enhancements of 5.8 and 9.8, respectively.

2 Related Work

2.1 Instance Segmentation

Two-stage methods. Two-stage instance segmentation methods achieve state-of-the-art performance by detecting bounding boxes and then performing segmentation in each region of interest (RoI). FCIS [19] introduces the position-sensitive score maps within an instance proposal for mask segmentation. Mask R-CNN [9] incorporates Faster R-CNN [24] with an additional FCN branch to perform segmentation within the detected box. To strengthen feature representation, PANet [22] further integrates multi-level features of FPN. With the guidance of spatial contexts, HTC [2] proposes an effective cascade framework with multi-level box and mask heads. DynaMask [17] proposes mask selection module with adaptive resolution.
One-stage methods. One-stage instance segmentation methods directly regress segmentation masks without the need for bounding box detection. PolarMask [32] builds a polar coordinate and performs instance segmentation by regressing dense distance between contour points and the polar centre. YOLACT [1] proposes to model instance masks with a weighted combination of different prototypes. SOLO [27] directly generates instance masks with the guidance of "instance categories". E2EC [35] proposes a dynamic matching strategy for accurate contour generation. SharpContour [36] applies a learnable contour initialization architecture to improve the quality of the boundary. Compared with two-stage solutions, these methods have simpler mechanisms but less accurate results.
Boundary-aware segmentation. The methods mentioned above primarily focus on improving network architectures for better recognition and segmentation accuracy. In contrast, recent works in instance segmentation aim to improve boundary accuracy, assuming that the backbone networks for feature extraction are already good enough. PolyTransform [20] generates segmentation results in polygons and then refines them for accurate boundaries. SegFix [33] introduces the concept of "interior pixels" to fix inaccurate boundary segmentation results. PointRend [15] uses point-based predictions iteratively in blurred areas for high-quality segmentation. BMask R-CNN [5] explicitly incorporates a boundary segmentation loss to improve boundary segmentation. RefineMask [34] optimizes instance boundaries by focusing on boundary regions in later stages of refinement. Transfiner [13] introduces a transformer-based refinement module to enhance the segmentation of boundary regions.

While boundary-aware segmentation methods have made notable advancements, they primarily focus on inter-mask similarity within instances. In contrast, BEIS stands out by exploring both intra-mask and cross-image dissimilarity. This strategy guides the network to generate distinct features across instance boundaries. Additionally, BEIS seamlessly integrates into existing two-stage solutions with minimal extra computational overhead.

2.2 Contrastive Learning

Contrastive learninghas recently made significant advancements in classification and object detection. SimCLR [3] introduces unsupervised contrastive learning to obtain high-quality representations. In SimCLR, positive pairs are formed from features of different views of the same image, and negative pairs are generated from features of different images. MoCo [10] introduces a feature bank with a momentum encoder to reduce the memory cost and increase the number of negative keys. BYOL [7] presents an asymmetric framework consisting of target and online networks to remove the negative keys.

In the application of segmentation, SIOD [16] proposes a pixel-level group contrastive learning to mine latent instances from feature representation space. [25] utilizes contrastive learning to perform object discovery for object detection. Moreover, targeting a specific segmentation problem, DenseCL [28] performs dense contrastive learning at the level of pixels. CBL [26] performs category-based decision boundary separation in local neighborhoods. Besides, Some previous methods methods [14, 18, 23] propose to perform supervised contrastive learning to utilize class labels to prevent image features with the same category from wrongly detaching. Recently, ContrastMask [29] uses pixel-level contrastive learning to mine instances from unseen classes.

In contrast, our BEIS framework focuses on holistic contrastive perspective to enhance boundary discrimination for instance segmentation. We introduce adaptive pair construction strategies to tackle complex boundary scenarios while minimizing computational overhead. This makes BEIS unique compared to other methods.

3 Boundary-Enhanced Instance Segmentation

3.1 Overview

The framework proposed for Boundary-Enhanced Instance Segmentation (BEIS) is illustrated in Figure 3. It builds upon the established two-stage Mask R-CNN architecture, introducing an additional *boundary-enhanced feature* incorporating two efficient *boundary discrimination losses*. Given an RoI feature map and its corresponding semantic feature maps from the FPN, a feature fusion module is employed. This module, comprising 3 convolution layers,

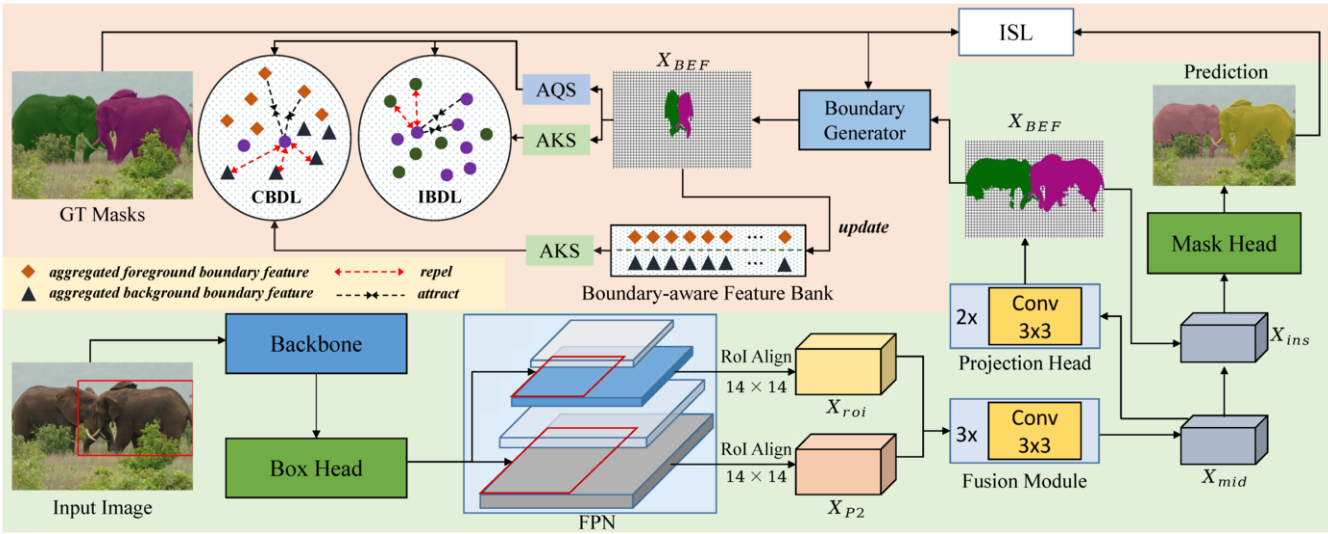

Figure 3. Overview of the Boundary-Enhanced Instance Segmentation (BEIS) framework. Only components in the green region are used during testing. IBDL and CBDL refer to intra-mask and cross-image boundary discrimination learning respectively. AQS and AKS indicate adaptive query and key feature selection. ISL represents the standard instance segmentation loss. The boundary-aware feature bank stores aggregated features from X_{BEF} and supplies key features for CBDL. Backbone and Box Head details are omitted for conciseness.

	Object Region		Boundary Region	
	WM	*CI*	*WM*	*CI*
baseline	0.346	0.129	0.720	0.375
+ CBDL	0.339	0.127	0.607	0.268
+ IBDL	0.325	0.120	0.541	0.297

Table 1. Comparison of mean pixel-level cosine similarity (CS) *within-mask (WM)* and *cross-image (CI)* of segmentation features on COCO *val* set. Both CBDL and IBDL substantially alleviate feature confusion across boundaries.

generates an enhanced feature map labeled as X_{mid}. Subsequently, the boundary-enhanced feature X_{BEF} is extracted from X_{mid} using a projection head employing two 3×3 convolutions. Supervised by the Intra-mask Boundary Discrimination Loss (IBDL) and the Cross-image Boundary Discrimination Loss (CBDL), X_{BEF} models boundary distinguishability. Finally, X_{BEF} is concatenated with X_{mid}, yielding segmentation features X_{ins} for generating final segmentation masks.

Next, we will define the Boundary Region, introduce the proposed Boundary-Enhanced Training involving IBDL and CBDL, and present BEIS along with a multi-stage refinement pipeline.

3.2 Definition of the Boundary Region

Consider $M_k \in R^{H_k \times W_k}$, representing the k-th instance mask, with H_k and W_k denoting its height and width. The boundary region of M_k is the set of pixels adjacent to contour pixels. We define a boundary pixel set B_k for this region. The formulation is as follows:

$$B_k = \{x_i | x_i \in M_k, d_i \leq t\}, \tag{1}$$

where d_i signifies the Euclidean distance from pixel x_i to the nearest mask contour pixel, and $t = 3$ is a predefined threshold. To optimize efficiency, we use a convolution operator for approximate calculations. B_k consists of two subsets, b_1 and b_0, which each specifically encompass the foreground and background pixels, correspondingly.

3.3 Boundary-Enhanced Training

Boundary segmentation has been an enduring challenge in the instance segmentation, as boundary regions often confuse networks due to semantic similarities. Prior research has mainly focused on re-weighting boundary segmentation [5] and intricate post-processing [15, 13, 34] to refine boundaries. Yet, the challenge of distinguishing boundary features has been neglected. As demonstrated in Table 1, features across boundaries, within the same mask or different images, display notably higher similarity compared to non-boundary features. This high similarity leads to inaccurate boundary segmentation by deep networks. Consequently, separating boundary features is crucial for enhancing boundary segmentation. In this section, we present our approach of intra-mask and cross-image boundary discrimination learning on X_{BEF}, using fully-supervised pixel losses to resolve boundary mask conflicts.

3.3.1 Intra-mask BDL

We introduce the intra-mask boundary discrimination loss L_{ibdl} to cultivate discriminative boundary-aware features within the boundary region B_k, leveraging the boundary-enhanced features X_{BEF} as follows:

$$L_{ibdl} = \sum_{b_i \subset B_k} \sum_{b_j \subset B_k, i \neq j} L_{ibdl}^{i,j}, \tag{2}$$

$$L_{ibdl}^{i,j} = \frac{1}{n_{b_i}} \sum_{x_l \in b_i} -\log \frac{\sum_{x_m \in b_j} e^{(-D(f_l, f_m))}}{\sum_{x_n \in (b_i \cup b_j)} e^{(-D(f_l, f_n))}}, \tag{3}$$

where f_l represents the mask feature of pixel x_l, and n_{b_i} is the count of boundary pixels in b_i. $D(.,.)$ denotes the L_2 distance. Discriminative targets effectively segregate mask features in different instances' boundary regions, enhancing segmentation accuracy. For discriminating b_i with b_j, we select 16 query features from b_i, and for each query, 4 positive and 32 negative keys are chosen from $(b_i \cup b_j)$.

3.3.2 Cross-image BDL

Intra-mask boundary discrimination learning improves boundary segmentation by exploring boundary relationships within masks. However, real-world scenarios have a more complex distribution of boundary scenes than the training data. Instances can encounter diverse backgrounds, necessitating a broader focus. Thus, solely relying on intra-mask relations may not capture this diversity sufficiently.

Figure 3 illustrates the proposed boundary-aware feature bank T, enhancing cross-image boundary discrimination learning. This feature bank comprises two sub-banks, T_0 and T_1, designated for storing background and foreground boundary features, respectively. To diversify the stored embedding, we introduce aggregated representations of either the foreground or background boundary from a single mask into the feature bank. The aggregated boundary feature R_k^c is defined as $R_k^c = \frac{\sum_{x_i \in B_k} \mathbb{I}(l_i=c) f_i}{\sum_{x_i \in B_k} \mathbb{I}(l_i=c)}$, with $\mathbb{I}(.)$ indicating foreground ($c = 1$) or background ($c = 0$). The bank solely storing aggregated features incurs negligible memory usage. The cross-image discrimination loss is then directly formulated as:

$$L_{cbdl} = \frac{-1}{|B_k|} \sum_{x_i \in B_k} \log \frac{\sum_{R_j \in T_l, l \neq l_i} e^{(-D(f_i, R_j))}}{\sum_{R_s \in (T_0 \cup T_1)} e^{(-D(f_i, R_s))}}. \quad (4)$$

CBDL efficiently uses negative samples from the feature bank to enhance holistic learning and complement IBDL. In implementation, 16 query features are sampled per mask, with 16 positive keys and 128 negative keys selected from the feature bank for each query. We implement the boundary feature bank using a first-in, first-out queue with a length of 65,536. The memory cost is calculated as 2 (two banks) × 65536 × 256 (feature dims) × 4 (number of bytes occupied by float32) ÷ 1048576 (number of bytes in one MB) = 128 MB.

3.3.3 Adaptive Query and Key Feature Selection

Employing pixel-level discrimination learning across numerous boundary regions within multiple instance masks entails significant computational demands and substantial memory resources. To address this, our study introduces an adaptive query and key feature selection algorithm, focusing on the most discriminative pairs.

Adaptive Query Selection. Segmentation networks adeptly distinguish common boundaries prevalent in training data. Hence, we choose to sample challenging queries with low confidences below a defined threshold $\delta = 0.97$:

$$B_{fore}^{hard} = \{x_i | x_i \in B_k, y_i < \delta, l_i = 1\}, \quad (5)$$

$$B_{back}^{hard} = \{x_i | x_i \in B_k, y_i > 1 - \delta, l_i = 0\}, \quad (6)$$

where y_i denotes the predicted confidence of x_i post Sigmoid normalization. Adaptive query selection focuses on boundary pixels, categorized as background (B_{back}^{hard}) or foreground (B_{fore}^{hard}).

Adaptive Key Selection. In instance segmentation, networks effectively distinguish between different instances (e.g., person and bird). However, challenges arise when delineating boundaries between instances with close semantic relationships (e.g., two people). Randomly sampling negative keys for each query, without considering semantic confusion, proves suboptimal. To address this, we propose a non-uniform negative key selection strategy. We establish a category confusion matrix P to define probabilities for selecting boundary pixels of different categories as negative keys. $P(u, v)$ denotes the probability of selecting boundary pixels of instances with category v as negative keys, when the query instance belongs to category u. We initialize P with 0.1 and update it during training based on feature confusion:

$$p(u, v) = (R_k^u \cdot R_k^v), \forall v \in C_{M_k}, \quad (7)$$

$$P(u, v) = \lambda * P(u, v) + (1 - \lambda) * p(u, v). \quad (8)$$

R_k^u and R_k^v are aggregated boundary features of the query instance with category u in mask M_k (with category set C_{M_k}) and other instances with category v, respectively. $(\cdot)$ calculates the normalized cosine similarity. An exponential moving average (EMA) strategy with coefficient $\lambda = 0.99$ is proposed to smooth $P(u, v)$.

Negative keys for IBDL and CBDL are dynamically chosen based on P. In IBDL, for query of class u, the probability of selecting a negative key with class v in M_k is $\frac{\exp(P(u,v))}{\sum_{w \in C_{M_k}} \exp(P(u,w))}$. In CBDL, for query of class u, the probability of selecting a negative key with class v in the boundary feature bank is $\frac{\exp(P(u,v))}{\sum_{w \in C} \exp(P(u,w))}$, where C is the dataset's category set. Adaptive negative key selection prioritizes confusing pairs, aiding the segmentation network in improving boundary discrimination.

3.4 Multi-stage Refinement and Loss Function

While the proposed BEIS framework significantly enhances boundary segmentation, the ultimate performance is constrained by the 28×28 mask resolution. BEIS effectively enhances boundary discrimination and harmoniously integrates with various mask segmentation resolutions, like 112×112. To unleash BEIS's potential, we introduce a multi-stage refinement instance segmentation framework.

Initially, a 14×14 coarse instance mask is generated using standard Mask R-CNN and a fusion module consisting of three 3×3 convolutional layers. Subsequently, a multi-stage refinement process iteratively elevates mask quality. Each stage takes three inputs: instance features from the prior stage, instance mask, and boundary-enhanced features X_{BEF}. These inputs are combined using three 3×3 convolutional layers, and the resulting features are upscaled. The mask head reiterates the refinement process, culminating in a high-quality instance mask up to 112×112. Within each refinement module, an independent projection head produces X_{BEF} based on instance features X_{ins}, followed by BEIS. Loss functions for initial mask prediction and subsequent refinement stages are:

$$L_{BEIS} = L_{box} + \sum_{i=1}^{4} (L_{mask,i} + (L_{ibdl,i} + L_{cbdl,i})). \quad (9)$$

L_{box} is box detection loss. $L_{mask,i}$, $L_{ibdl,i}$, and $L_{cbdl,i}$ correspond to mask segmentation, intra-mask and cross-image discrimination learning losses at the i-th stage, respectively. Formulations for L_{box} and L_{mask} are defined in Mask R-CNN. The BEIS exhibits robustness when it comes to loss weight, and all loss weights are uniformly assigned a value of 1.0.

4 Experiments

4.1 Datasets and Implementation Details

Our experimentation covers COCO, LVIS, and Cityscapes, widely-used benchmarks for instance segmentation evaluation. COCO holds $118k$ training, $5k$ validation, and $20k$ test images in the $train2017$

subset. LVIS comprises 2 million annotations across $1,203$ categories, with $100k$ training, $20k$ validation, and $20k$ test images. Cityscapes, tailored for autonomous driving, features $2,975$ training, 500 validation, and $1,525$ high-resolution (2048×1024) test images. BEIS is also evaluated on the COCO-OCC val-set [12], housing $1,005$ images with densely-overlapping objects.

The implementation of BEIS is based on Mask R-CNN. Training schedules for the backbone, box head, and mask head followed Detectron2's [31] standards. We evaluated BEIS's components on COCO val2017 through comprehensive ablation studies. In ablation studies, unless otherwise specified, BEIS was trained with R50-FPN using a $1\times$ learning schedule.

4.2 Ablation Experiments

Different numbers of negative keys. Analyzing the influence of negative keys on IBDL and CBDL performance (Table 2 and 3), models with more negative keys exhibit improved results for both. However, excessive negative keys provide marginal gains while increasing memory demands during training. Thus, a balanced number of negative keys is chosen for IBDL and CBDL.

Boundary-Enhanced feature. Investigating the impact of the boundary-enhanced feature X_{BEF} (Table 4), fusion feature X_{BEF} from X_{roi} and X_{P2} outperforms using just X_{roi} or semantic feature maps from $P2$ in FPN. Moreover, results suggest that BEIS benefits from finer feature maps like $P2$.

Multi-stage BEIS. Comparing multi-stage refinement (Table 5), with all models trained using $3\times$ schedules, the one-stage BEIS with 28×28 masks competes with SOTA, with minor runtime impact (15.0 FPS) compared to Mask R-CNN. Both the two-stage and three-stage BEIS enhance accuracy with limited computational cost. Three-stage BEIS (112×112 masks) achieves a balance between precision and complexity, making it the preferred choice.

Adaptive Query and Key Selection. Table 6 shows adaptive query and key strategies provide more improvements than treating all pairs equally. This underscores the value of hard negative keys with indistinguishable semantics in boundary segmentation, as demonstrated by the 1.3 AP_B increase. As shown in Table 7, adaptive query selection is not sensitive to the value of δ.

Intra-mask and cross-image discrimination learning. In analyzing the impact of IBDL and CBDL (Table 8), with models trained with $3\times$ schedules, IBDL boosts AP by 1.3 and boundary AP_B by 1.4. CBDL alone enhances AP by 0.8 and boundary AP_B by 1.0. Combining both achieves further improvements. Both techniques add little to computational load, being used only in training.

N	AP_{val}	AP_B	AP_B^{50}	AP*
8	37.0	24.4	49.3	40.1
16	37.2	24.5	49.4	40.9
32	**37.5**	**24.7**	**49.7**	**41.2**
64	37.3	24.5	49.5	41.0

Table 2. Numbers of negative keys for IBDL.

4.3 Comparison with State-of-the-Arts

We present comparisons between our approach against state-of-the-art methods on COCO, Cityscapes and LVIS-1.0. Evaluation metrics encompass AP_{val}, boundary AP_B [4], AP*, and AP_{test}.

N	AP_{val}	AP_B	AP_B^{50}	AP*
32	37.2	24.4	49.4	40.7
64	37.4	**24.7**	49.4	40.8
128	**37.5**	**24.7**	**49.7**	**41.2**
256	37.4	24.5	49.6	41.0

Table 3. Numbers of negative keys for CBDL.

Feature	AP_{val}	AP_B	AP_B^{50}	AP*
X_{P2}	36.8	23.7	48.3	40.4
X_{roi}	37.3	24.0	49.2	40.7
$X_{P2} + X_{roi}$	**37.5**	**24.7**	**49.7**	**41.2**

Table 4. Effect of boundary-enhanced features (BEF).

Stage	Output Size	AP_{val}	AP_B	AP*	FPS
1	28×28	39.3	25.9	42.7	15.0
2	56×56	39.8	26.4	43.5	13.5
3	112×112	40.5	**27.0**	**44.7**	12.3
4	224×224	**40.6**	26.8	44.3	8.4

Table 5. Mask AP obtained using multi-stage BEIS.

AQS	AKS	AP_{val}	AP_B	AP*
		36.7	23.4	40.1
✓		37.0	24.0	40.6
	✓	37.3	24.4	41.0
✓	✓	**37.5**	**24.7**	**41.2**

Table 6. Effectiveness of AQS and AKS.

δ	AP_{val}	AP^B	AP_{50}^B	AP*
0.99	37.3	24.6	49.6	41.1
0.97	**37.5**	**24.7**	**49.7**	**41.2**
0.95	37.4	24.6	49.4	41.2

Table 7. Effect of δ in AQS.

IBDL	CBDL	AP_{val}	AP_B	AP_B^{50}	AP*
		38.9	25.2	47.5	42.0
✓		40.2	26.6	49.4	44.3
	✓	39.7	26.2	49.0	43.8
✓	✓	**40.5**	**27.0**	**49.7**	**44.7**

Table 8. Effectiveness of IBDL and CBDL.

COCO. Table 9 compares BEIS to COCO's SOTA instance segmentation methods. Across various backbones, BEIS consistently outperforms PatchDCT and DynaMask by 1.6 AP and 1.1 AP with an R101-FPN. Remarkably, even against multi-stage refinement methods like Mask Transfiner and RefineMask that employ additional semantic segmentation and boundary loss optimizations, BEIS, using an R50-FPN, outperforms Mask R-CNN and PointRend

Method	Backbone	Resolution	AP_{val}	AP_{test}	AP^*	AP_B	AP_S	AP_M	AP_L	FPS
Mask R-CNN (ICCV17)	R50-FPN	28×28	37.2	37.3	38.2	21.2	18.6	39.5	53.3	15.7
BMask R-CNN (ECCV20)	R50-FPN	28×28	38.2	38.4	41.7	25.2	19.2	39.8	**54.1**	12.4
BCNet (CVPR21)	R50-FPN	28×28	38.4	38.6	41.4	25.0	21.9	40.9	49.3	9.4
PointRend (CVPR20)	R50-FPN	224×224	38.5	38.6	41.9	25.7	19.1	40.6	53.8	11.0
Mask Transfiner (CVPR22)	R50-FPN	112×112	38.8	39.3	41.9	25.8	21.6	40.8	49.8	6.2
BEIS (Ours)	R50-FPN	112×112	**40.5**	**40.9**	**44.7**	**27.0**	**23.1**	**43.6**	53.9	12.3
Mask R-CNN (ICCV17)	R101-FPN	28×28	38.6	39.0	40.8	23.1	19.5	41.4	50.5	13.4
BCNet (CVPR21)	R101-FPN	28×28	39.8	40.2	42.8	26.1	22.7	42.4	51.1	7.5
HTC (CVPR19)	R101-FPN	112×112	40.2	40.5	43.2	26.4	21.9	42.6	56.3	4.7
PointRend (CVPR20)	R101-FPN	224×224	40.1	40.4	43.6	26.7	23.2	42.9	54.6	9.7
Mask Transfiner (CVPR22)	R101-FPN	112×112	40.0	40.5	43.0	26.9	22.5	42.2	53.6	5.5
RefineMask (ICCV21)	R101-FPN	112×112	40.8	41.2	44.6	27.8	23.4	44.6	55.0	11.0
PatchDCT (ICLR23)	R101-FPN	112×112	40.5	40.7	45.0	27.6	20.8	43.3	**57.7**	11.4
DynaMask (CVPR23)	R101-FPN	112×112	41.0	41.3	44.6	27.4	22.8	43.4	54.0	8.3
BEIS (Ours)	R101-FPN	112×112	**42.1**	**42.3**	**45.7**	**29.2**	**25.0**	**45.4**	55.4	11.2
Mask Transfiner (CVPR22)	Swin-B	112×112	44.9	44.5	45.4	28.2	27.8	47.6	58.5	3.5
PatchDCT (ICLR23)	Swin-B	112×112	46.6	46.4	47.8	29.4	29.0	49.0	**59.9**	7.3
BEIS (Ours)	Swin-B	112×112	**47.8**	**47.6**	**49.0**	**31.4**	**31.5**	**51.6**	59.4	7.4

Table 9. Comparison with SOTA methods. All methods are trained on COCO $train2017$ with ImageNet-pretrained weights and $3\times$ schedules. AP_{val}, AP_{test}, and AP^* are evaluated on COCO-val, COCO-$test$, and COCO-val sets with LVIS annotations. FPS is measured on one single V100.

by 3.3 AP and 2.0 AP, respectively. Notably, BEIS also significantly improves boundary AP_B over these methods, reaffirming its effectiveness in boundary discrimination. Additionally, experiments on the LVIS-0.5 val-set, with more accurate boundary annotations, demonstrate further improvement. Importantly, BEIS's efficiency stands out among multi-stage refinement methods since most discrimination learning modules are used only during training. These outcomes highlight BEIS's capacity to generate high-quality masks.

Cityscapes. Table 10 displays the Cityscapes benchmark results. BEIS achieves the best mask AP of 39.2 and boundary AP_B of 21.2. Compared to the baseline Mask R-CNN, BEIS substantially enhances boundary AP_B from 11.4 to 21.2, evidencing the effectiveness of the boundary enhancement framework. Furthermore, BEIS surpasses other state-of-the-art (SOTA) methods like Mask Transfiner and PatchDCT by margins of 4.8 and 2.4 AP_B, respectively, with an R-50 FPN model and ImageNet-pretrained weights. These significant improvements, leveraging high-quality annotations in Cityscapes, underscore BEIS's superiority. All models are trained with 64 epochs and R50-FPN for fair comparisons.

LVIS-1.0. Table 11 presents LVIS-1.0 dataset results, where BEIS achieves the highest mask AP of 26.5, surpassing the baseline by 3.2 AP_f, showcasing its efficiency. All Models are trained with 1× schedule for fair comparisons.

COCO-OCC. BEIS's performance is compared against other occlusion-aware segmentation models (BCNet, and MS R-CNN [11]) using the COCO-OCC split. Table 12 demonstrates BEIS outperforming BCNet by 3.4 AP and 5.1 AP_B, indicating its proficiency in boundary segmentation, particularly for occlusion scenarios. All Models are trained with 1× schedule for fair comparisons.

Qualitative Results. Figure 4 presents qualitative comparisons on COCO, demonstrating that our BEIS achieves notably higher mask precision and quality, particularly in overlapping regions. In Figure 5, we present qualitative comparisons on the Cityscapes dataset. Our BEIS framework, using 112×112 mask size, produces more accurate and precise predictions compared to RefineMask with 112×112 mask size, and PointRend with a 224×224 mask size.

Method	Resolution	AP_B	AP_B^{50}	AP_{val}	AP_{50}
Mask R-CNN (ICCV17)	28×28	11.4	37.4	33.8	61.5
BMask R-CNN (ECCV20)	28×28	15.6	45.7	36.2	62.6
DCTMask (CVPR21)	112×112	14.6	44.5	36.9	62.9
PointRend (CVPR20)	224×224	16.6	47.2	35.9	61.8
RefineMask (ICCV21)	112×112	18.0	50.2	37.6	63.3
Mask Transfiner (CVPR22)	112×112	16.4	46.0	36.0	62.1
PatchDCT (ICLR23)	112×112	18.8	51.0	38.2	64.5
DynaMask (CVPR23)	112×112	18.4	49.5	38.0	63.6
BEIS (Ours)	112×112	**21.2**	**53.6**	**39.2**	**65.3**

Table 10. Performance comparison on Cityscapes val-set.

Method	AP_{val}	AP_r	AP_c	AP_f
Mask R-CNN (ICCV17)	22.1	10.1	21.7	30.0
RefineMask (ICCV21)	25.5	**14.2**	24.3	31.7
BEIS (Ours)	**26.5**	13.5	**25.1**	**33.2**

Table 11. Performance comparison on LVIS-1.0 val set.

Method	AP_{val}	AP_B	AP_{50}
Mask R-CNN (ICCV17)	29.7	13.7	49.9
MS R-CNN (CVPR19)	30.3	16.2	50.0
BCNet (CVPR21)	31.7	17.4	51.1
BEIS (Ours)	**35.1**	**22.5**	**55.8**

Table 12. Performance comparison on COCO-OCC val-set.

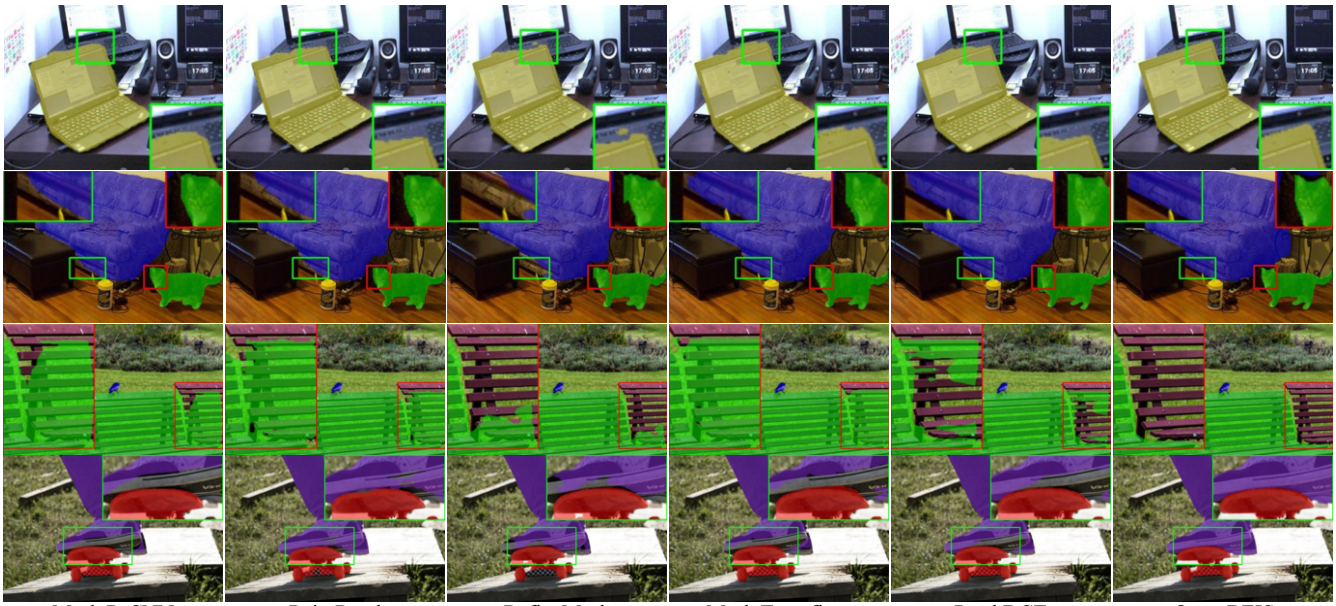

Figure 4. Qualitative comparisons with Mask R-CNN, PointRend, Mask Transfiner, RefineMask, and PatchDCT on COCO.

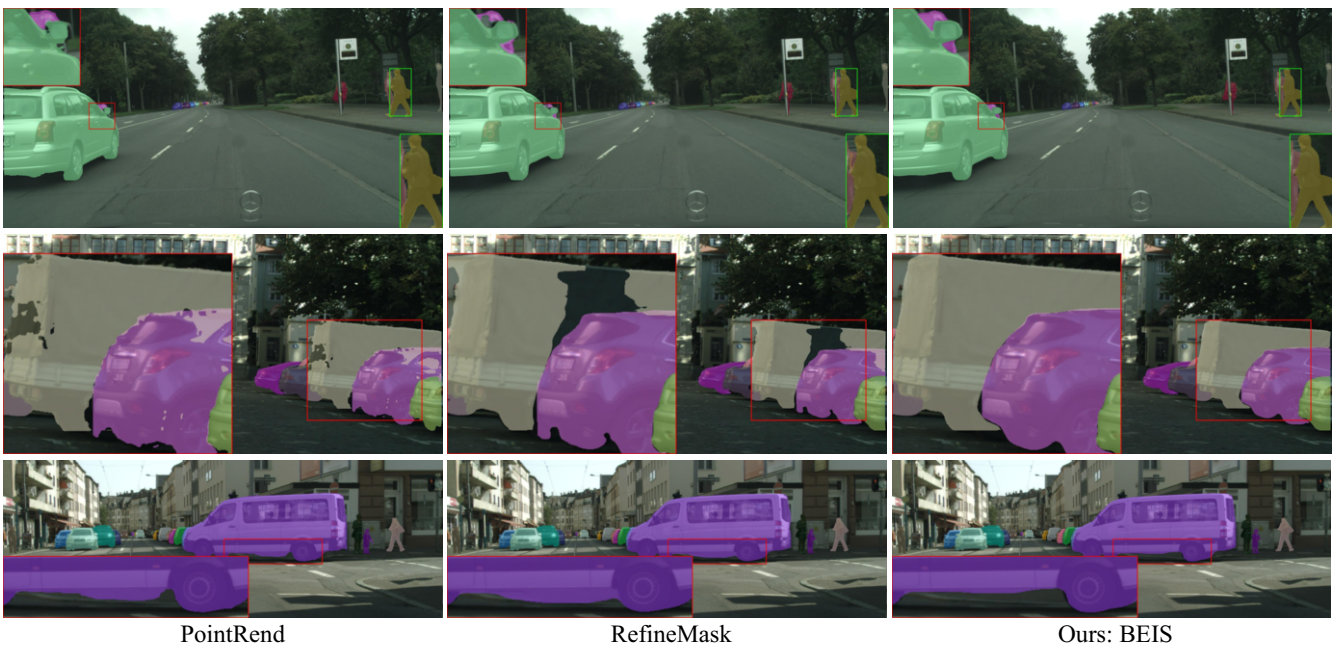

Figure 5. Qualitative comparisons with instance segmentation methods PointRend and RefineMask on Cityscapes.

5 Limitation and Conclusion

In this study, we introduce BEIS, a novel and effective instance segmentation method that prioritizes boundary enhancement. BEIS employs intra-mask and cross-image boundary-aware discrimination learning to augment the discernibility of boundary regions, resulting in accurate boundary and overall segmentation. Unlike previous approaches focusing solely on intra-mask features, BEIS incurs minimal computational and memory overhead while yielding substantial segmentation improvements. Our experiments on COCO, Cityscapes, and LVIS benchmarks demonstrate BEIS's superiority. However, the current efficacy of BEIS is constrained by the quality of segmentation annotations. Future efforts will target this issue by refining annotations based on predictions. Meanwhile, we will also apply BEIS for video and point cloud instance segmentation.

Acknowledgements

This work was supported by the NSFC under Grant 62072271.

References

[1] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee. YOLACT: real-time instance segmentation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[2] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, C. C. Loy, and D. Lin. Hybrid task cascade for instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

[4] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] T. Cheng, X. Wang, L. Huang, and W. Liu. Boundary-preserving mask R-CNN. In *European Conference on Computer Vision (ECCV)*, 2020.

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[7] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[8] A. Gupta, P. Dollár, and R. B. Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask R-CNN. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[10] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[11] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang. Mask scoring R-CNN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[12] L. Ke, Y. Tai, and C. Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[13] L. Ke, M. Danelljan, X. Li, Y. Tai, C. Tang, and F. Yu. Mask transfiner for high-quality instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[14] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[15] A. Kirillov, Y. Wu, K. He, and R. B. Girshick. Pointrend: Image segmentation as rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[16] H. Li, X. Pan, K. Yan, F. Tang, and W. Zheng. SIOD: single instance annotated per category per image for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[17] R. Li, C. He, S. Li, Y. Zhang, and L. Zhang. Dynamask: Dynamic mask selection for instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[18] S. Li, X. Xia, S. Ge, and T. Liu. Selective-supervised contrastive learning with noisy labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[20] J. Liang, N. Homayounfar, W. Ma, Y. Xiong, R. Hu, and R. Urtasun. Polytransform: Deep polygon transformer for instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[21] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[22] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[23] S. Liu, S. Zhi, E. Johns, and A. J. Davison. Bootstrapping semantic segmentation with regional contrast. In *International Conference on Learning Representations (ICLR)*, 2022.

[24] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[25] J. Seo, W. Bae, D. J. Sutherland, J. Noh, and D. Kim. Object discovery via contrastive learning for weakly supervised object detection. In *European Conference on Computer Vision (ECCV)*, 2022.

[26] L. Tang, Y. Zhan, Z. Chen, B. Yu, and D. Tao. Contrastive boundary learning for point cloud segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[27] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li. SOLO: segmenting objects by locations. In *European Conference on Computer Vision (ECCV)*, 2020.

[28] X. Wang, R. Zhang, C. Shen, T. Kong, and L. Li. Dense contrastive learning for self-supervised visual pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[29] X. Wang, K. Zhao, R. Zhang, S. Ding, Y. Wang, and W. Shen. Contrastmask: Contrastive learning to segment every thing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[30] Q. Wen, J. Yang, X. Yang, and K. Liang. Patchdct: Patch refinement for high quality instance segmentation. *International Conference on Learning Representations (ICLR)*, 2023.

[31] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[32] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo. Polarmask: Single shot instance segmentation with polar representation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[33] Y. Yuan, J. Xie, X. Chen, and J. Wang. Segfix: Model-agnostic boundary refinement for segmentation. In *European Conference on Computer Vision (ECCV)*, 2020.

[34] G. Zhang, X. Lu, J. Tan, J. Li, Z. Zhang, Q. Li, and X. Hu. Refinemask: Towards high-quality instance segmentation with fine-grained features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[35] T. Zhang, S. Wei, and S. Ji. E2EC: an end-to-end contour-based method for high-quality high-speed instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[36] C. Zhu, X. Zhang, Y. Li, L. Qiu, K. Han, and X. Han. Sharpcontour: A contour-based boundary refinement approach for efficient and accurate instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.