2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

BEE-Net: Bridging Semantic and Instance with Gated Encoding and Edge Constraint for Efficient Panoptic Segmentation

Xinyang Huang[†,1,3], Guanghui Zhang[†,1], Dongchen Zhu[1,3], Yunpeng Sun[2], Wenjun Shi[1], Gang Ye[2], Yang Xiao[2], Lei Wang[1,3], Xiaolin Zhang[1,3,4], Bo Li[2], and Jiamao Li[1,3,*]

***Abstract*— Panoptic segmentation is a challenging perception task, which can help robots to comprehensively perceive the surrounding environment. In the task, we notice that semantic, instance, and panoptic have rich relations, however, which are rarely explored. In this work, we propose a novel panoptic, instance, and semantic bridged network to delve into the reciprocal relation. To make semantic and instance benefit from each other, we design a novel Gated Encoding (GE) module, incorporating complementary cues between semantic and instance heads through the gated mechanism. In addition, a novel edge-aware consistency constraint among edges of each task is presented, which exhaustedly exploits geometric constraints, to boost the segmentation quality of challenging edges. Experimental results on the Cityscapes and MS-COCO datasets demonstrate that our approach achieves state-of-the-art performance in an efficient CNN-based paradigm, attaining a balance between accuracy and efficiency.**

I. INTRODUCTION

Panoptic segmentation aims to assign each pixel a semantic label and each object a unique instance ID [1]. It offers a more comprehensive understanding of scenes, facilitating intelligent scene perception for robots, and it also has broad application prospects in autonomous driving [2] and medical image analysis [3], [4].

Recent years have witnessed rapid development in panoptic segmentation due to the urgency of comprehensive understanding. The existing models could be roughly divided into CNN-based and Transformer-based paradigms according to network structure [5]. The former segments scenes by utilizing multi-scale feature extraction and capturing local patterns [6], [8], [29], which gains promising segmentation accuracy at a low computational burden. The latter improves segmentation accuracy through serialized encoding, but usually at the cost of efficiency, leading to poor practicability in the field of robotics. In this work, we will follow the CNN-based paradigm to support the practicality of robots.

Most existing panoptic segmentation methods generally delve into semantic and instance tasks [8], [15], [30], however, rarely fully explore the reciprocal relation among semantic, instance, and panoptic. Intuitively, pixels of different classes must belong to different instances, while pixels of the same instance must belong to the same category. Moreover, the logical OR of the edges of semantic and instance predictions should match the panoptic edge. This means that these tasks contained in panoptic segmentation could cooperate with each other by seeking common grounds while reserving differences, leading to potential mutual assistance theoretically.

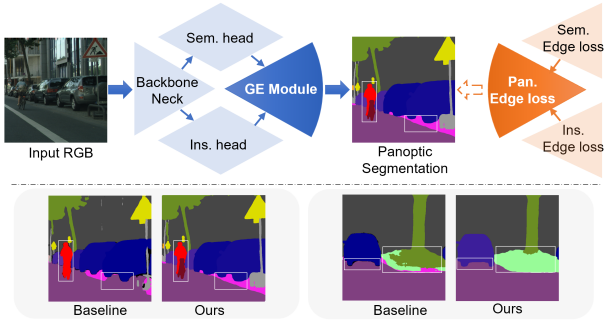

Fig. 1. An illustration of the proposed BEE-Net, which explores the reciprocal association from two aspects: associate gated encoding corresponding to GE module and panoptic-instance-semantic geometric edge consistency constraint. BEE-Net produces more discriminative predictions for objects and more accurate predictions at challenging edges (shown with white boxes). The baseline is EfficientPS [29].

Currently, some algorithms have begun to explore the work of multi-task assistance [7], [8], [30]. AUNet [7] conducts only unidirectional assistance, failing to fully utilize the potential of bidirectional benefits. Further, BANet [30] proposes a bidirectional learning method to utilize local information from each other. However, it ignores global contextual information. When encountering complex scenes (*e.g.* multi-object scene), the performance of these methods will probably decrease due to insufficient exploration of reciprocity association or consistency constraints. How to deeply mine the the reciprocal relations from the perspectives of global feature assistance and coupling constraints is promising to alleviate this problem. Besides, to the best of our knowledge, there is no prior work exploring the global coupling constraint among panoptic, instance, and semantic to improve panoptic prediction.

In this paper, we propose a panoptic-instance-semantic bridged efficient panoptic segmentation network via gated encoding and edge constraint, named BEE-Net. Our BEE-Net delves into the reciprocal association from two aspects: associate gated encoding and pan-ins-sem edge-aware

†These authors contributed equally to this work.
*National Science and Technology Major Project from Minister of Science and Technology, China(2018AAA0103100), Shanghai Sailing Program (23YF1456300), Youth Innovation Promotion Association, Chinese Academy of Sciences(2021233, 2023242), Shanghai Academic Research Leader(22XD1424500)
[1]Bionic Vision System Laboratory, State Key Laboratory of Transducer Technology, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China. *Corresponding author: Jiamao Li jmli@mail.sim.ac.cn)
[2]Lotus Robotics Ltd.
[3]University of Chinese Academy of Sciences, Beijing 100049, China.
[4]ShanghaiTech University, Shanghai 201210, China.

consistency constraint. The pan-ins-sem refers to panoptic-instance-semantic for simplification. Figure 1 illustrates the framework and shows that BEE-Net performs better than the baseline. Specifically, we first present a novel lightweight Gated Encode (GE) module that efficiently bridges the semantic and instance branches, leveraging the complementary between the two tasks. The module benefits from both the structural 'thing' class cue advantages of the instance branch and pixel-level context cue advantages of the semantic branch, to better distinguish instances and boost category recognition performance. In addition, based on the finding that the logic OR between instance and semantic edges matches the panoptic edge, we propose a geometric edge-aware consistency loss. In this way, the part of edges among panoptic, instance, and semantic should be aligned, explicitly constraining the optimization direction between tasks to a certain extent. To reduce the influence of slight geometric transformation at edge, an instance-oriented Inverseform loss is constructed, finally, improving the segmentation accuracy of panoptic edge. We evaluate the proposed BEE-Net on Cityscapes [11] and MS-COCO [12] datasets. The results show that our model not only exhibits outstanding segmentation accuracy but also demonstrates remarkable efficiency. In summary, the contributions of this work are as follows:

- We propose a novel panoptic-instance-semantic bridged panoptic segmentation network, which effectively delves into the reciprocal relation among the tasks via global gated encoding and edge consistency constraint.
- We design a lightweight gated encoding (GE) module to embrace global complementary cues between semantic and instance heads, allowing the two heads to benefit from each other.
- We are the first to propose explicit panoptic-instance-semantic edge-aware consistency loss to incorporate consistency constraints among the tasks, further boosting the segmentation quality of challenging edges.
- Experiments on the Cityscapes and MS-COCO datasets demonstrate the effectiveness of our approach, and the proposed BEE-Net achieves state-of-the-art accuracy, exhibiting a trade-off between accuracy and efficiency.

II. RELATED WORK

A. CNN-Based Panoptic Segmentation

In recent years, the task of image segmentation has benefited from the rapid development of Convolutional Neural Networks (CNN). Similarly, panoptic segmentation, as a fusion of semantic and instance segmentation, has also seen advancements due to this progress. Fully supervised panoptic segmentation models based on CNN mainly include top-down models, bottom-up models, and other types of models [5]. The top-down models generally perform semantic and instance segmentation separately and then fuse the results. For instance, models such as UPSNet [8], Panoptic FPN [13], and JSIS-Net [14] segment 'stuff' and 'thing' categories individually and subsequently fuse them to produce panoptic segmentation results. In the bottom-up models, semantic segmentation is performed first, and then instance masks are generated by classification and refinement to generate panoptic segmentation results. A representative of this model type is Panoptic-DeepLab [15], with Axial-Deeplab [16] and PanoNet [17] also following the same idea. Other types of CNN-based panoptic segmentation models include K-Net [18], which performs segmentation via a set of learnable convolution kernels, etc. In general, the panoptic segmentation model based on CNN has good segmentation accuracy and remarkable inference speed. It is worth noting that, the top-down model can also be optimized for classic segmentation issues such as occlusion and edge delineation by incorporating new methods before the fusion stage.

B. Transformer-based Panoptic Segmentation

With the introduction of the attention mechanism, the transformer model was quickly applied to machine translation and other fields. Subsequently, the encoder-decoder structure was introduced into the field of computer vision [19], [20]. The field of panoptic segmentation is no exception. Max-Deeplab [21] directly predicts the category masks through the transformer structure without bounding boxes or object center prediction. CMT [22] combines transformer and clustering approaches to achieve panoptic results. The latest Mask2former [23] performs mask prediction and category prediction through the mask-attention mechanism. Generally, transformer-based panoptic segmentation models offer high accuracy. However, due to complexity and numerous parameters, they consume significant memory and have slower training speeds. Its inference speed is somewhat lacking compared to the CNN-based panoptic segmentation models.

III. METHOD

As shown in Figure 2, our BEE-Net is composed of the shared backbone and neck, as well as two task-specific parallel heads. One head is for pixel-level semantic prediction, while the other is for handling the object-level instance segmentation problem. Similar to EfficientPS [29], multi-scale features are obtained through 2-way FPN, and then the semantic and instance heads learn the useful information for separate tasks via corresponding semantic and instance decoders, respectively. Between the two heads, the proposed GE module is introduced to make them cooperate with each other, thereby boosting their representation capability. Finally, the panoptic segmentation result is obtained through fusing semantic and instance predictions like [29]. To strengthen the perception of challenging edges, we exploit the proposed edge-aware consistency loss (E-Loss) to consider consistency among panoptic, semantic, and instance domains while emphasizing edge geometric transformation consistency. The E-loss is combined with other classic losses in Mask-RCNN[10] to supervise the training of the network.

A. GE Module

Although there are differences between instance and semantic segmentation tasks, there exists a correlation between

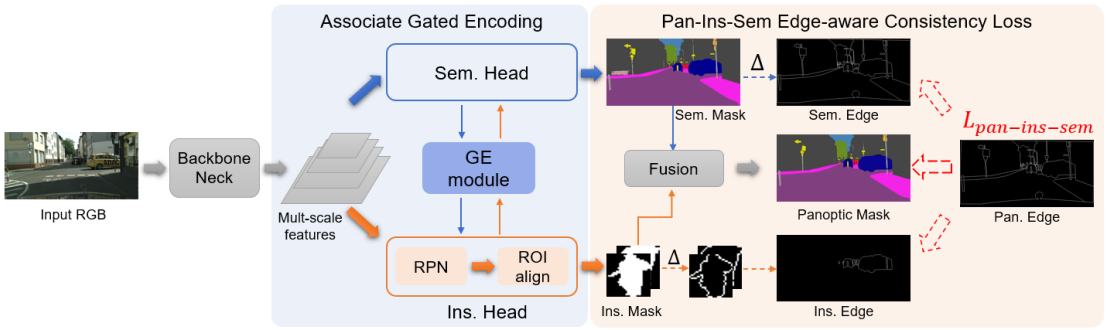

Fig. 2. Overview of BEE-Net. The input is a single RGB image and the output is a panoptic mask, which is generated by fusing the semantic and instance predictions. Multi-scale features from the backbone and neck are fed into the semantic and instance heads, respectively. The GE module establishes an association between the two heads, and the edge-aware consistency loss (E-Loss) is used for jointly optimizing the edge prediction, further boosting the quality of panoptic segmentation. Δ denotes edge extraction with Laplace Operator. RPN and ROI align denote Region Proposal Network and Region of Interest Alignment [10].

the two. For example, pixels with different semantics must belong to different instances. Due to the oversight of this correlation, existing methods generally suffer from sub-optimal accuracy. Figure 1 provides an example, the semantic head predicts a person and a bicycle, while the instance head identifies them as one instance, causing the inconsistency problem of "pixels with different semantics belong to an instance", and this ultimately leads to incorrect predictions. To alleviate this problem, we propose a novel GE module to incorporate the correlation, as shown in Figure 3.

The inputs F_{sem} and F_{ins} are multi-scale features obtained by preliminary abstraction in the semantic and instance heads. For the instance head, instances belonging to the 'thing' category will be recognized, which are relatively fine-grained, local, and structured. We first try to enhance semantic information by incorporating structural 'thing' information, producing instance-aware semantic features F_{i-sem}. The procedure can be expressed as:

$$F_{i-sem} = f_2[f_1(F_{ins}) \otimes F_{sem}] \tag{1}$$

where f_1 aims to summarize key foreground 'thing' features, and f_2 is committed to fusing instance and original semantic features. $\otimes$ indicates element-wise multiplication. This makes foreground 'thing' and background 'stuff' more discriminative, which is conducive to later panoptic fusion.

Subsequently, based on enhanced semantic features F_{i-sem}, similar operations are implemented to realize the semantic-aware instance feature F_{s-ins}, as formulated in Eq. 2. This process enables the instance features to summarize the global semantic context, improving its perception of object semantics and the surrounding environment, thereby enriching the instance expressiveness.

$$F_{s-ins} = g_2[g_1(F_{i-sem}) \otimes F_{ins}] \tag{2}$$

where the functions of g_1 and g_2 are similar to f_1 and f_2 in Eq. 1, respectively.

Specifically, taking the acquisition process of instance-aware semantic feature as an example, the f_1 is implemented by employing a composite convolution of three layers followed by a sigmoid activation to generate the weights, and then it is element-wise multiplied with F_{sem} to obtain weighted feature termed as F_{wsem}. This provides the F_{sem} with 'thing' features. This process can be formulated as:

$$f_1 = \sigma\{[ReLU(BN(Conv(x_1)))]^{[3]}\} \tag{3}$$

where $Conv$ denotes a 3×3 convolution. σ, BN, and $ReLU$ stand for sigmoid, batch normalization, and ReLU activation. x_1 denotes the input F_{ins} here.

Inspired by residual learning, we fuse the weighted features and original features through element-wise sum to ensure the preservation of integral information. Then, drawing on the function of channel attention [31], we recalibrate the fused feature via compound operation including GAP, Conv, and BN, to ultimately boost the discrimination of semantic features. This process corresponds to f_2 (Eq. 4).

$$f_2 = \sigma[BN(Conv(GAP(F_{isem})] \otimes [x_2 \oplus F_{sem}] \tag{4}$$

where GAP denotes global average pooling. $\oplus$ indicates element-wise sum. x_2 refers to the weighted feature F_{wsem} here.

For the enhancement of instance features using semantic features, symmetric implementation is employed.

To be clear, although our GE module and BANet [30] both belong to the category of bidirectional feature mutual assistance. However, they have fundamental distinctions. The BANet only interacts with local features at corresponding locations primarily, ignoring contextual information. In contrast, our GE module perceives global knowledge and incorporates the relations between the object and background as well as scene-level perception. Furthermore, thanks to the application of GAP, which significantly compresses the feature resolution, our module is relatively lightweight. The inference time in Table I in experiments confirms that our method demands almost no extra time compared to the baseline.

B. Edge-aware Consistency Loss

The panoptic segmentation at the edges has always been a challenge. The exploration of reciprocity relations among edges may be promising to alleviate this problem. From Figure 4(a), we notice that the logical OR between instance and semantic edges should match the panoptic edge, which

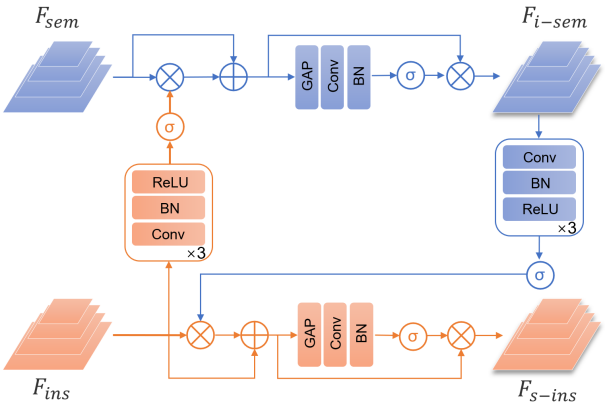

Fig. 3. The structure of Gated Encoding Module (GE).

is simple yet effective. However, it is ignored in existing works. Hence, we propose the edge-aware consistency loss. The constraint can be described as:

$$(P_{sem}||P_{ins}) - T_{pan} \to 0, \tag{5}$$

where $P_{sem}||P_{ins}$ is written as P_{pan} for simplification. $||$ refers to logical OR. P_{sem} and P_{ins} refer to the predicted semantic and instance edges, respectively. T_{pan} denotes the ground truth of the panoptic edge.

Instance edge P_{ins} is obtained by merging edges of multiple instance masks resized to the original image size using Roi-sampling. Notably, the multiple masks selected here are the masks with top n confidence.

$$P_{ins} = \sum_{i=1}^{n} Roisampling(BMask_i), \tag{6}$$

where $BMask_i$ denotes the edge of the instance mask output. $Roisampling$ refers to the operation that restores the mask to its original size.

For the supervision to P_{pan}, the binary cross-entropy loss (L_{bce}) may be the most straightforward approach. However, considering that many edges (*i.e.*, P_{sem}, P_{ins}, and P_{pan}) are involved in the supervision, as well as semantic and instance edges are obtained independently. When a task prediction undergoes small geometric transformations, the consistency constraint among edges may be disrupted, leading to unstable convergence. As a result, L_{bce} might be insufficient to handle such geometry transformation, as shown in Figure 4(b).

Inspired by the Inverseform[9] for semantic segmentation, we introduce the Inverseform loss (L_{if}, see Eq. 7) to supervised the P_{sem} and P_{pan}. For object-level P_{ins}, different from the uniform and regular division manner in [9], to retain the integrity of the instances, we propose an instance-oriented edge geometric transformation consistency loss. Specifically, we utilize the output 28×28 masks in the instance head (see Figure 2) to build the Inverseform loss.

$$L_{if}(P,T) = FC([P,T]), \tag{7}$$

where FC denotes the pre-trained fully connected layer that is capable of measuring edge geometry transformation[9]. P and T denote the prediction and ground truth, respectively.

In summary, to achieve consistency constraint, P_{pan}, P_{ins}, and P_{ins} are supervised via a weighted combination of L_{bce} and L_{if}, respectively. The specific formulas are as follows:

$$L_e = \alpha L_{if}(P,T) + \beta L_{bce}(P,T), \tag{8}$$

where α , β are the weight factors. P corresponds to P_{pan}, P_{ins}, and P_{ins}. T corresponds to the T_{pan}, T_{ins}, and T_{ins}. L_e is consist of L_{e-sem}, L_{e-ins}, L_{e-pan}.

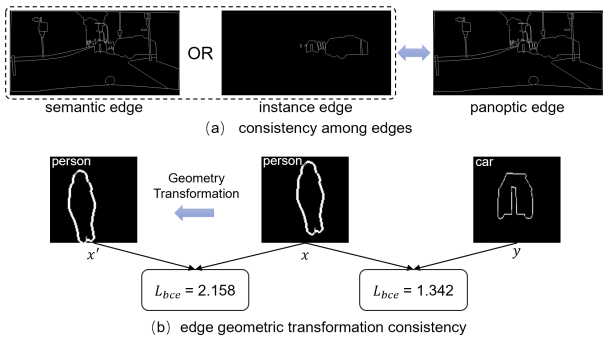

Fig. 4. The illustration of edge-aware consistency loss (E-Loss). L_{bce} fails for geometry transformations of edges. x and y are edges of mask predictions. x' is generated by only applying a mild shift to x.

C. Training Strategy

The total loss L comprises L_{sem}, L_{rpn}, L_{rcnn}, L_{mask}, and L_e, representing the losses of semantic segmentation, RPN, RCNN, mask in [10], and our E-Loss respectively.

$$L = \lambda_1 L_{sem} + \lambda_2 L_{rpn} + \lambda_3 L_{rcnn} + \lambda_4 L_{mask} + \lambda_5 L_e. \tag{9}$$

IV. Experiments

In this section, we evaluate the proposed BBE-Net on Cityscapes[11] and MS-COCO[12] datasets, achieving SOTA accuracy and speed compared with other methods.

A. Datasets and Metrics

Cityscapes The dataset has 2975 images for training and 500 images for validation with fine annotations. We report experimental results on val set with 19 semantic labels and 8 annotated instance categories.

MS-COCO The dataset consists of 115k images for training and 5k images for validation. The panoptic annotations include 80 'thing' categories and 53 'stuff' categories.

Evaluation Metrics We adopted the most commonly used and classic evaluation metrics PQ (panoptic quality), SQ (segmentation quality), and RQ (recognition quality) [1] to evaluate our method. Metrics with superscripts 'th' and 'st' (*i.e.*, PQ_{th}, PQ_{st}) denote 'thing' or 'stuff' classes performance respectively. In addition, we also evaluate semantic segmentation via mIoU metric.

Implementation Details We trained our network on a single A100 GPU using EfficientNetB5 as the backbone and 2-way FPN as the neck. In L_e, $\{\alpha, \beta\}$ are set to $\{0.3, 0.3\}$, $\{0.2, 0.1\}$, and $\{0.1, 0.1\}$ when processing semantic, instance, and panoptic edges, respectively. Hyperparameters λ_1 to λ_7 were set to $\{1.0,1.0,1.0,0.3,1.0\}$ for MS-COCO and $\{0.7,1.0,0.5,1.0,1.0\}$ for Cityscapes. For Cityscapes, we employ the SGD optimizer, setting the learning rate, momentum, weight decay, warmup iters, and warmup ratio to 0.07, 0.9, 0.0001, 500, and 0.3 respectively, with a batch size

TABLE I

QUANTITATIVE EVALUATION ON THE CITYSCAPES VAL SET. WE COMPARE OUR METHOD WITH OTHER STATE-OF-THE-ART METHODS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD. PANOPTIC-FCN[24], UPSNET[8], PANOPTIC-DEEPLAB[15], AND EFFICIENTPS[29] ARE CNN-BASED METHODS, WHILE COPS[25], MASK2FORMER[26], AND CMT[22] ARE TRANSFORMER-BASED METHODS.

Method	PQ	SQ	RQ	PQ_{th}	PQ_{st}	mIoU	inference time(s)
Panoptic-FCN[24]	61.4	-	-	54.7	66.3	-	-
UPSNet[8]	61.8	81.3	74.8	57.6	64.8	79.2	-
EfficientPS[29]	63.3	81.2	76.7	58.8	66.6	-	0.15
Panoptic-Deeplab[15]	64.1	-	-	-	-	81.5	-
COPS[25]	62.1	-	-	55.1	67.2	-	0.19
Mask2Former[26]	62.4	-	-	55.7	67.3	-	-
CMT[22]	64.6	**82.6**	77.4	-	-	81.4	0.20
Ours	**65.0**	81.9	**78.4**	**59.7**	**68.8**	**81.6**	**0.15**

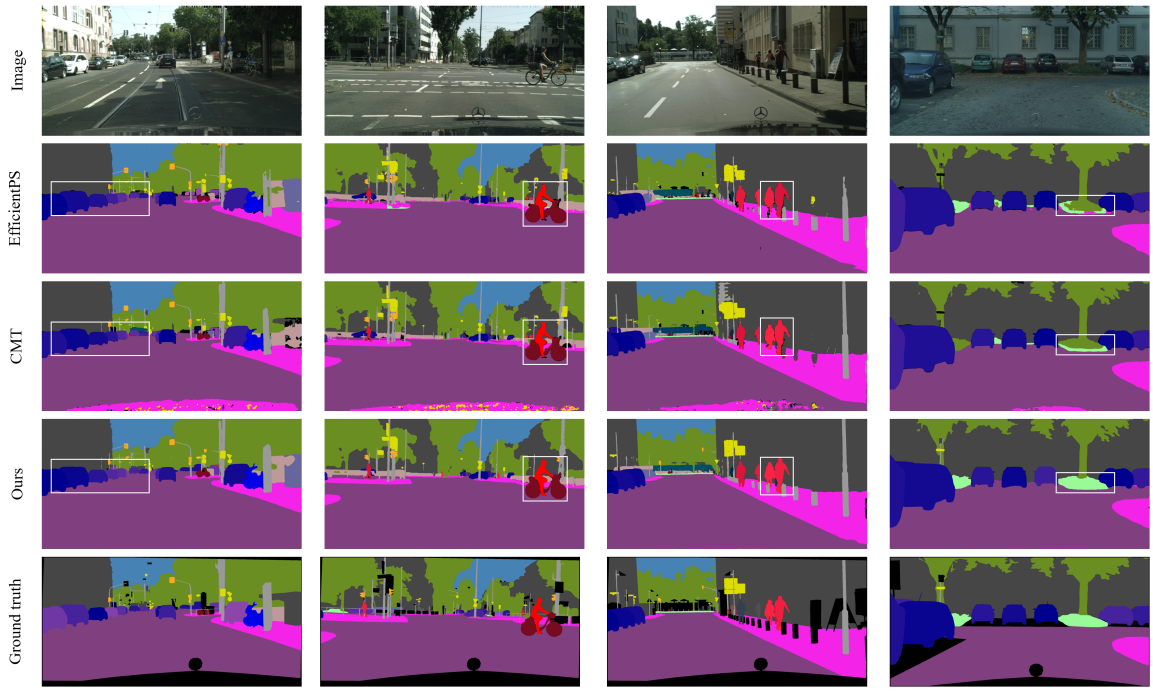

Fig. 5. Visualized results of BEE-Net on Cityscapes val set (%). We compare our method with EfficientPS[29] and CMT[22]. It is evident that we have made significant improvements in both instance category recognition and edge optimization.

of 8. For MS-COCO, we employ the Adam optimizer, setting the learning rate, weight decay, warmup iters, and warmup ratio to 0.0002, 0.0001, 30000, and 0.9, with a batch size of 30. The images from both datasets were randomly flipped and scaled by a factor of 0.5 to 2.0×. Images from Cityscapes are cropped up to 1024×2048 pixels, and those from MS-COCO up to 960×960 pixels. Notably, we train our models without extra data on both datasets.

B. Comparison with State-of-the-Art Methods

Results on Cityscapes As illustrated in Table I, our proposed model achieves a remarkable PQ performance of 65.0% without any extra training data or label bank. Compared to the baseline EfficientPS [29], our method outperforms it by 1.7% PQ, and achieves 0.9% and 2.2% improvements in categories 'thing' and 'stuff', respectively. In terms of inference time (s) on the image with 1024×2048 pixels, it is worth mentioning that our model is significantly superior to Transformer-based methods. Our BBE-Net realizes an approximate 20% improvement in inference speed with slightly higher accuracy over CMT [22]. Notably, compared with EfficientPS, our model almost has no time cost increase, thanks to our lightweight GE module.

Figure 5 visualizes the panoptic segmentation results. We can observe that our model achieves more refined predictions, especially for edges, which indicates the proposed E-Loss plays a significant role in edge prediction. Simultaneously, thanks to our GE module, different instances are better distinguished, and the mispredictions in the category have been rectified. In conclusion, our model demonstrates superior accuracy coupled with commendable efficiency in the panoptic segmentation task.

Results on MS-COCO In Table II, we compare our method with other popular SOTA approaches on MS-COCO dataset. EfficientPS[29] serves as the baseline that we meticulously replicated and optimized for our study. CBT [27] is a recently proposed CNN-base method, whereas [28] and [21] represent new Transformer-based models. It can be observed that our method achieves PQ 51.6%, which is superior to baseline EfficientPS by 2.7% in a large margin and outperforms

TABLE II

QUANTITATIVE EVALUATION ON THE MS-COCO VAL SET (%). WE COMPARE OUR METHOD WITH OTHER STATE-OF-THE-ART METHODS. CBT[27] AND EFFICIENTPS[29] ARE CNN-BASED METHODS, WHILE PANOPTIC-SEGFORMER AND MAX-DEEPLAB-L[21] ARE TRANSFORMER-BASED METHODS.

Method	PQ	SQ	RQ
CBT[27]	42.9	78.1	52.8
EfficientPS[29]	48.9	78.8	60.7
Panoptic-SegFormer[28]	50.6	-	-
MaX-DeepLab-L[21]	51.1	-	-
Ours	**51.6**	**79.2**	**63.8**

Fig. 6. A close-up ablation study example of GE module and E-Loss. Using the GE module clearly distinguishes the person and bicycle, and the E-Loss further improves the edge segmentation quality significantly.

Transformer-based models [28] and [21]. Furthermore, experimental results on this dataset indicate that our method is applicable to a broader range of arbitrary scenarios.

C. Ablation Study

From Table III, it can be seen that incorporating either our GE module or E-Loss individually performs better than the baseline model in all metrics. Owing to the synergistic integration from both feature and edge perspectives, our model achieves optimal accuracy, with a PQ improvement of 1.7% over baseline.

TABLE III

ABLATION STUDY FOR GE MODULE AND E-LOSS ON CITYSCAPES VAL SET (%).

Method	GE	E-Loss	PQ	SQ	RQ
Baseline[29]			63.3	81.2	76.7
Ours	✓		63.7	81.3	77.4
		✓	64.6	81.8	78.0
	✓	✓	**65.0**	**81.9**	**78.4**

In addition to quantitative analysis, Figure 6 also presents a qualitative example. Comparing the baseline with the model incorporating only the GE module, the enhanced feature-level consistency constraints introduced by the GE module facilitate effective segmentation of both 'person' and 'bicycle'. Further contrasting the model with only GE against the model integrating both GE and E-Loss, one can observe more correct predictions at the edges, which demonstrates the effectiveness of E-Loss on improving edges.

Ablation Study on GE module

Results of the ablation study on GE module are reported in Table IV. In the 'order' column, 'S⇒I' denotes first enhancing semantic features and then instance features, and 'I-Sem' denotes enhancing semantic features with instance information. We conduct an analysis of unidirectional assistance as well as the sequential interplay in the context of bidirectional assistance scenarios. Experimental results indicate that 'S⇒I' exhibits a marginal superiority over 'I⇒S'. This might be attributed to the fact that the instance 'thing' information is relatively local and structured, while the semantic information encompasses both 'thing' and 'stuff' categories. Correspondingly, this means that instance predictions might be more correct in the early stage than semantic predictions.

TABLE IV

ABALTION STUDY FOR GE MODULE SETTINGS ON CITYSCAPES VAL SET (%).

Method	Order	S⇒I	I⇒S	PQ	SQ	RQ
Baseline[29]	-	-	-	63.3	81.2	76.7
Ours	-	✓	-	63.5	81.2	77.1
	-	-	✓	63.4	81.2	77.0
	I1S2	✓	✓	63.6	81.0	77.4
	S1I2	✓	✓	**63.7**	**81.3**	**77.4**

Ablation Study on Edge-aware Consistency Loss

We perform detailed ablation experiments on our proposed edge-aware consistency loss, as presented in Table V. Here, 'L_{if}' represents the incorporation of L_{e-sem}&L_{e-ins} within the independent semantic and instance branches, and 'L_{e-pan}' denotes pan-ins-sem consistency among edges, including L_{if} and L_{bce}. It is clear from the results that employing either module individually boosts segmentation accuracy. And the combination of both modules leads to the best results of 1.3% PQ improvements. Furthermore, as observed in Figure 5 and 6, our approach significantly improves the quality of edge segmentation.

TABLE V

ABALTION STUDY FOR THE E-LOSS ON CITYSCAPES VAL SET (%).

Method	L_{e-sem}&L_{e-ins}	L_{e-pan}	PQ	SQ	RQ
Baseline[29]			63.3	81.3	76.7
Ours	✓		64.4	81.5	77.9
		✓	64.2	81.4	77.7
	✓	✓	**64.6**	**81.8**	**78.0**

V. CONCLUSIONS

We propose a novel efficient panoptic segmentation model BEE-Net, which bridges semantic and instance with gated encoding and edge constraint. the gated encoding module (GE) aims to leverage the complementary information between semantic and instance to enhance the quality of panoptic segmentation. From the perspective of the loss constraint, the proposed edge-aware consistency loss (E-Loss) delves into the relation among panoptic, semantic, and instance edges, significantly enhancing the segmentation accuracy of object boundaries. Benefiting from the utilization of reciprocal relations, final experimental results indicate that our method achieves state-of-the-art performance while striking a balance between accuracy and efficiency.

REFERENCES

[1] Kirillov, Alexander, et al. "Panoptic segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[2] Petrovai, Andra, and Sergiu Nedevschi. “Multi-task network for panoptic segmentation in automated driving.” 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019.
[3] Liu, Dongnan, et al. “Panoptic feature fusion net: a novel instance segmentation paradigm for biomedical and biological images.” IEEE Transactions on Image Processing 30 (2021): 2045-2059.
[4] Zhang, Donghao, et al. “Panoptic segmentation with an end-to-end cell R-CNN for pathology image analysis.” Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. Springer International Publishing, 2018.
[5] Li, Xinye, and Ding Chen. “A survey on deep learning-based panoptic segmentation.” Digital Signal Processing 120 (2022): 103283.
[6] Liu, Dongnan, et al. “Nuclei Segmentation via a Deep Panoptic Model with Semantic Feature Fusion.” IJCAI. 2019.
[7] Li, Yanwei, et al. “Attention-guided unified network for panoptic segmentation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[8] Xiong, Yuwen, et al. “Upsnet: A unified panoptic segmentation network.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.
[9] Borse, Shubhankar, et al. “Inverseform: A loss function for structured boundary-aware segmentation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[10] He, Kaiming, et al. “Mask r-cnn.” Proceedings of the IEEE international conference on computer vision. 2017.
[11] Cordts, Marius, et al. “The cityscapes dataset for semantic urban scene understanding.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[12] Lin, Tsung-Yi, et al. “Microsoft coco: Common objects in context.” Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer International Publishing, 2014.
[13] Kirillov, Alexander, et al. “Panoptic feature pyramid networks.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[14] De Geus, Daan, Panagiotis Meletis, and Gijs Dubbelman. “Panoptic segmentation with a joint semantic and instance segmentation network.” arXiv preprint arXiv:1809.02110 (2018).
[15] Cheng, Bowen, et al. “Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[16] Wang, Huiyu, et al. “Axial-deeplab: Stand-alone axial-attention for panoptic segmentation.” European conference on computer vision. Cham: Springer International Publishing, 2020.
[17] Chen, Xia, Jianren Wang, and Martial Hebert. “PanoNet: Real-time panoptic segmentation through position-sensitive feature embedding.” arXiv preprint arXiv:2008.00192 (2020).
[18] Zhang, Wenwei, et al. “K-net: Towards unified image segmentation.” Advances in Neural Information Processing Systems 34 (2021): 10326-10338.
[19] Vaswani, Ashish, et al. “Attention is all you need.” Advances in neural information processing systems 30 (2017).
[20] Zhu, Xizhou, et al. “Deformable detr: Deformable transformers for end-to-end object detection.” arXiv preprint arXiv:2010.04159 (2020).
[21] Wang, Huiyu, et al. “Max-deeplab: End-to-end panoptic segmentation with mask transformers.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[22] Yu, Qihang, et al. “CMT: Clustering mask transformers for panoptic segmentation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[23] Cheng, Bowen, et al. “Masked-attention mask transformer for universal image segmentation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[24] Li, Yanwei, et al. “Fully convolutional networks for panoptic segmentation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[25] Abbas, Ahmed, and Paul Swoboda. “Combinatorial optimization for panoptic segmentation: A fully differentiable approach.” Advances in Neural Information Processing Systems 34 (2021): 15635-15649.
[26] de Geus, Daan, and Gijs Dubbelman. “Intra-batch supervision for panoptic segmentation on high-resolution images.” Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.
[27] Mao, Lin, et al. “CNN Based Transformer for Panoptic Segmentation.” Journal of Software (2022): 1-14.
[28] Li, Zhiqi, et al. “Panoptic segformer: Delving deeper into panoptic segmentation with transformers.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.
[29] Mohan, Rohit, and Abhinav Valada. “Efficientps: Efficient panoptic segmentation.” International Journal of Computer Vision 129.5 (2021): 1551-1579.
[30] Chen, Yifeng, et al. “Banet: Bidirectional aggregation network with occlusion handling for panoptic segmentation.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.
[31] Hu, Jie, Li Shen, and Gang Sun. ”Squeeze-and-excitation networks.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.