RFENet: Towards Reciprocal Feature Evolution for Glass Segmentation

Ke Fan[1] , **Changan Wang**[2] , **Yabiao Wang**[2] , **Chengjie Wang**[1,2] , **Ran Yi**[1*] and **Lizhuang Ma**[1*]
[1]Shanghai Jiao Tong University
[2]Tencent Youtu Lab
slipperyfrank@sjtu.edu.cn, {changanwang, caseywang, jasoncjwang}@tencent.com, ranyi@sjtu.edu.cn, ma-lz@cs.sjtu.edu.cn

Abstract

Glass-like objects are widespread in daily life but remain intractable to be segmented for most existing methods. The transparent property makes it difficult to be distinguished from background, while the tiny separation boundary further impedes the acquisition of their exact contour. In this paper, by revealing the key co-evolution demand of semantic and boundary learning, we propose a Selective Mutual Evolution (SME) module to enable the reciprocal feature learning between them. Then to exploit the global shape context, we propose a Structurally Attentive Refinement (SAR) module to conduct a fine-grained feature refinement for those ambiguous points around the boundary. Finally, to further utilize the multi-scale representation, we integrate the above two modules into a cascaded structure and then introduce a Reciprocal Feature Evolution Network (RFENet) for effective glass-like object segmentation. Extensive experiments demonstrate that our RFENet achieves state-of-the-art performance on three popular public datasets. Code is available at *https://github.com/VankouF/RFENet.*

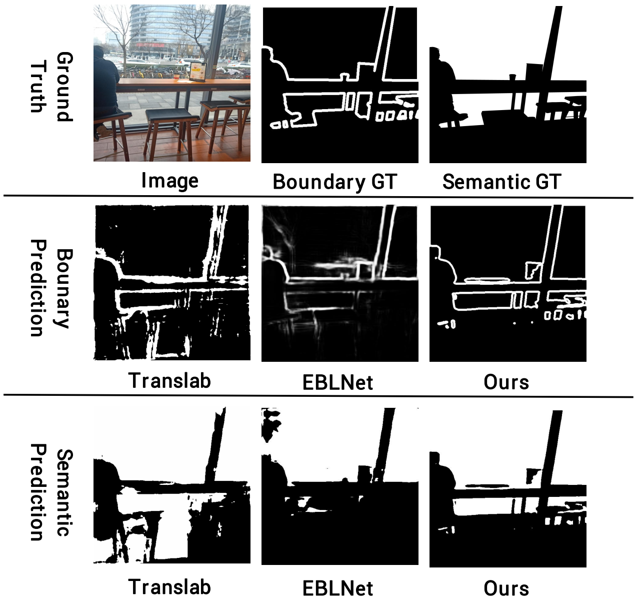

Figure 1: Illustration of our predicted boundary and semantic results, compared with the two state-of-the-art methods in glass-like object segmentation, *i.e.*, Translab and EBLNet. Our method produces a more accurate glass segmentation with a finer boundary, which is mainly contributed by our key feature co-evolution mechanism between semantic and boundary. The predictions of those ambiguous regions also gain the help from our SAR module.

1 Introduction

Detecting ubiquitous yet fragile glass-like objects is indispensable for vision based navigation systems. However, different from most other daily objects, glass-like objects are more confusing to be distinguished from background due to their transparent property. Besides, such objects mostly share an extremely thin separation boundary with background, making this task even more challenging.

Due to the above challenges, merely relying on either semantic content or separation boundary to segment out the glass regions is sub-optimal or at least inaccurate. Although glass segmentation methods mainly rely on semantic features to predict the semantic map, adding boundary information can still help to obtain global shape context, thus improving the performance of segmentation. Meanwhile, boundary prediction could also be improved by auxiliary semantic assistance, which can help to reduce the false edge prediction and in turn facilitate a more accurate glass segmentation. Being consistent with the above observation, some previous glass segmentation works explored either assistance from boundary to semantic or assistance from semantic to boundary: 1) Using boundary to assist semantic: [Xie *et al.*, 2020; Zheng *et al.*, 2022] introduced an auxiliary boundary supervision as a guidance to conduct glass segmentation refinement, which helps the prediction of those uncertain regions around the boundary; 2) Using semantic to assist boundary: [He *et al.*, 2021] proposed to supervise non-edge parts in a residual style to obtain finer edges and eliminate noisy edges in background, instead of directly performing a boundary feature enhancement. However, the above two paradigms,

[*]Corresponding authors.

i.e., using boundary to assist semantic, or using semantic to assist boundary, both ignored the importance of feature co-evolution from the two sides (*i.e.*, semantic branch and boundary branch). In other words, previous methods did not conduct the bi-directional assistance between the two branches, and simply adopting one-way assistance results in inferior performance.

To address these issues, in this paper, we propose an adaptive mutual learning mechanism to enable the *explicit* feature co-evolution between semantic branch and boundary branch. Such a mechanism helps to exploit the complementary information from each branch and is achieved by a novel Selective Mutual Evolution (SME) module. Specifically, the semantic feature is selectively enhanced with the guidance from the boundary branch, highlighting those weak response regions (especially for the pixels around boundary). In a similar way, the boundary feature is also selectively enhanced with the guidance from the semantic branch, mitigating the impact of edge noise from background or internal glass. With the above mutual learning strategy, the two branches reciprocally optimize each other to explore the intrinsic interdependence between them.

Despite the effectiveness of SME module, some regions remain indistinguishable, such as the pixels around boundary. To remedy this problem, we further propose a Structurally Attentive Refinement (SAR) module. To be more specific, we firstly sample a set of most reliable *boundary points* in the semantic feature, according to the confidence scores on the predicted boundary map, to capture the global shape information. Then a set of most *uncertain points* on glass segmentation map are enhanced with the semantic features of previous selected boundary points. Notably, this adaptive enhancement process is conditioned on the contents of those uncertain points, and is achieved with a cross-attention operation. In a nutshell, such an attentive feature refinement exploits extra boundary cues to help the inference of those ambiguous points, serving as a globally structural guidance. The proposed SAR module is pluggable with a simple design, and can also be applied to other boundary assisted methods.

Besides, inspired by pioneering exploration in the multi-scale feature representation, we further equip the above two modules with a cascaded style connection, benefiting from the progressive fusion of multiple receptive fields. Finally, we propose a Reciprocal Feature Evolution Network (RFENet) for glass-like object segmentation. We conduct extensive experiments against recent competitors on three popular glass-like object segmentation datasets and our RFENet achieves state-of-the-art performance. The visualized results in Figure 1 also demonstrates the superiority of our RFENet.

Overall, we summarize our contributions as follows:

- We propose RFENet, a novel glass-like object segmentation model, which achieves state-of-the-art performance on three popular benchmarks.
- We propose a Selective Mutual Evolution (SME) module to encourage the feature co-evolution of two branches, which effectively solves the inferior performance caused by only one-way assistance in the previous two-stream methods.
- We propose a Structurally Attentive Refinement (SAR) module to conduct further feature refinement for uncertain points with useful global shape context.

2 Related Work

Glass-like Object Segmentation. It is much more challenging to segment out glass-like objects than those common objects, mainly due to that inner glass regions often share extremely confusing appearance with surrounding background. To remedy this problem, some methods [Xu *et al.*, 2015; Chen *et al.*, 2018a; Huo *et al.*, 2022] resorted to exploit additional multi-modal information, such as 4D light-field, refractive flow map, and thermal image. Unfortunately, those multi-modal data is relatively expensive to acquire, which limits the wide applications. Instead, recent works [Yang *et al.*, 2019; Mei *et al.*, 2020; Lin *et al.*, 2021; Lin *et al.*, 2020; Xie *et al.*, 2020] contributed large-scale RGB image datasets for glass-like objects to promote research in related fields. However, due to the special property of glass-like objects, the off-the-shelf semantic segmentation methods[Chen *et al.*, 2018b; Zhao *et al.*, 2017] failed to achieve a promising performance. Similarly, many state-of-the-art salient object detection approaches [Pang *et al.*, 2020; Qin *et al.*, 2019; Zhuge *et al.*, 2022; Liu *et al.*, 2022] also result in an inferior prediction as the glass may not necessarily be salient.

Therefore, the demand for specialized methods recently attracts more attention in the field of glass-like object segmentation. [Yang *et al.*, 2019; Mei *et al.*, 2020; Lin *et al.*, 2021] tried to integrate abundant contextual or contrasted features to help distinguish glass regions, implying the importance of contextual information. [Ji *et al.*, 2023] utilizes a context and a texture encoder to extend the model from camouflaged object detection field into the glass detection area. Besides, [Xie *et al.*, 2020; He *et al.*, 2021; Lin *et al.*, 2020] proposed to segment glass-like objects under the assistance of boundary cues, benefiting from the high localization accuracy of boundary. Inspired by existing research, we further reveal the importance of feature co-evolution demand for glass segmentation and boundary learning. Based on this observation, we propose an adaptive mutual learning mechanism to effectively exploit the complementary information between semantic and boundary.

Boundary as Assistance. The boundary contour of glass objects clearly defines their distribution range with a pixel-level localization ability. As a result, the effective exploration of boundary cues becomes crucial in high-precision glass segmentation. Actually, previous research has also demonstrated that introducing boundary cues into their model plays an important role for semantic segmentation [Takikawa *et al.*, 2019; Ding *et al.*, 2019] and salient object detection [Wang *et al.*, 2019; Zhao *et al.*, 2019; Li *et al.*, 2018]. As for glass-like object segmentation, [Xie *et al.*, 2020; Zheng *et al.*, 2022] proposed to explicitly predict boundary map or decoupled boundary map, and utilized it as a guidance to assist the semantic stream. And [He *et al.*, 2021] tried to supervise edge part as well as none-edge part to explicitly model glass body and boundary. Both of them has proved the non-negligible performance gain brought by an appropriate

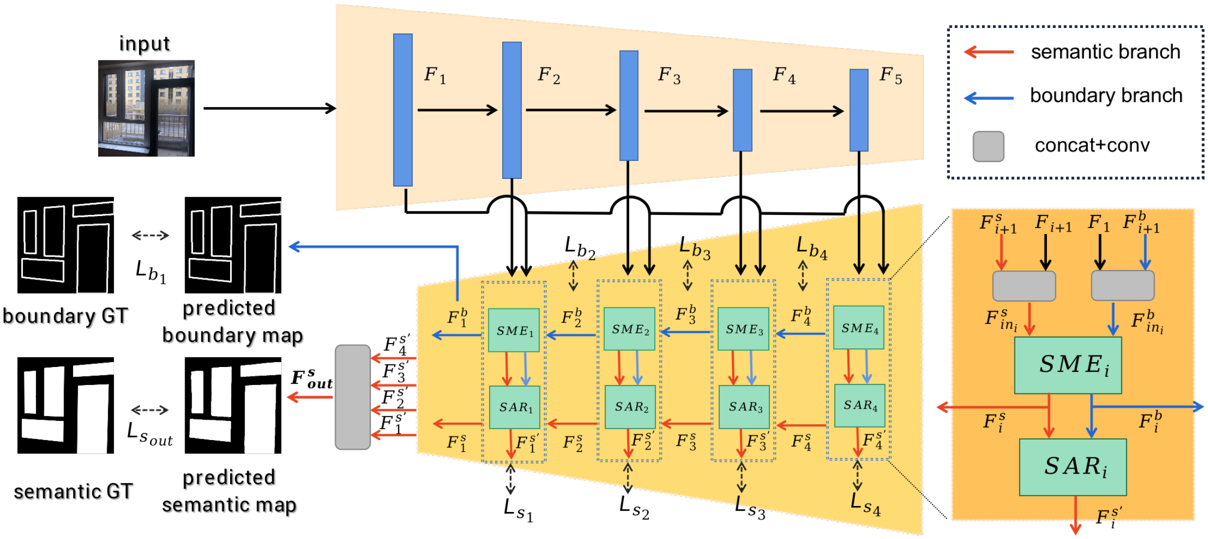

Figure 2: Overview of our proposed RFENet. The SME module enhances both semantic and boundary features in a reciprocal way, and the SAR module refines ambiguous points with global shape context. Both of them are combined repeatedly to form a cascaded structure.

exploit of boundary cues.

However, boundary prediction is often susceptible to background noise, especially for glass-like objects with transparent property. Different from existing methods, we propose an adaptive mutual learning mechanism to encourage feature co-evolution between semantic branch and boundary branch. Such a novel reciprocal structure helps to reduce the impact from background noise or internal glass reflection, with the guidance from semantic feature. Notably, the improved boundary prediction will in turns facilitate the glass segmentation in a cyclic enhancement style.

Glass Segmentation Refinement. The adoption of segmentation refinement technics has been demonstrated effective to further boost the model's prediction accuracy. For example, [Krähenbühl and Koltun, 2011; He *et al.*, 2010] proposed to use Conditional Random Fields and Guided Filter as post-processing to refine segmentation predictions. PointRend [Kirillov *et al.*, 2020] and MagNet [Huynh *et al.*, 2021] proposed to refine some selected point set using local or global context information. In the field of glass-like object segmentation, EBLNet [He *et al.*, 2021] proposed a PGM module to exploit global shape prior, which further improves the edge prediction precision.

Differently, we propose to adaptively aggregate useful shape context for the most ambiguous points instead of certain boundary points. And the refined point set is dynamically sampled along with the optimization process, imitating hard sample mining strategy. The proposed refinement module provides structural context for predictions of some local regions, such as boundary and reflective regions.

3 Method

3.1 Overview

The architecture of our RFENet is illustrated in Figure 2. There are two parallel branches in our RFENet: semantic branch and boundary branch. Specifically, we firstly adopt ResNet50 [He *et al.*, 2016] with ASPP module [Chen *et al.*, 2018b] as the backbone network to extract multi-scale deep features. Then Selective Mutual Evolution (SME) module is proposed to encourage the feature co-evolution of the two branches. Then we propose Structurally Attentive Refinement (SAR) module to further refine those uncertain points with global shape context. Finally, we integrate the two modules into a cascaded structure to exploit the internal hierarchical feature representation.

Formally, we denote the deep feature representations from different stages in backbone network as F_i, where $i \in \{1, 2, 3, 4, 5\}$ represents the i-th stage. The semantic branch predicts glass regions based on the last feature map F^s_{in} (*i.e.*, F_5), which contains the most rich context information. For the boundary map prediction, we use a concatenation[*] of F_1 and F_5 as input feature F^b_{in} to take advantage of the texture information from low-level features. As shown in Figure 2, SME takes semantic feature and boundary feature as inputs, and obtain the co-evolved features by a mutual operation. Then the evolved features are both fed into SAR module to conduct a further refinement for semantic features under the guidance of boundary cues. The final prediction is obtained by a sequentially stacking of the two modules.

During each stacking process, for SME, we fuse lower-level features to recover the textural details for semantic branch, which gradually aggregates useful multi-scale context information. For boundary prediction, we repeatedly fuse the finest feature F_1 into the input feature of every stage to exploit more detailed texture information. The whole stacking process within SME module could be formulated as:

[*] We perform bilinear interpolation on F_5 so that it has the same resolution as F_1, *i.e.*, 1/4 of the original image size. We use the same operation for other features F_i when necessary.

$$F_i^s, F_i^b = \begin{cases} \mathrm{SME}_i\left(F_{i+1}, [F_{i+1}; F_1]\right), & \text{if } \; i = 4, \\ \mathrm{SME}_i\left(\left[F_{i+1}^s; F_{i+1}\right], \left[F_{i+1}^b; F_1\right]\right), & \text{else.} \end{cases} \tag{1}$$

where $[\cdot]$ represents the feature concatenation operation, F_i^s and F_i^b represent the co-evolved semantic and boundary features output by SME. Then F_i^s and F_i^b are input into the SAR module and obtain a refined semantic feature $F_i^{s'}$ under the guidance of F_i^b, which could be formulated as:

$$F_i^{s'} = \mathrm{SAR}\left(F_i^s, F_i^b\right). \tag{2}$$

The concatenation of $F_i^{s'}$, $i \in \{1, 2, 3, 4\}$ is used as the final semantic feature F_{out}^s, which is responsible for the prediction of final semantic map. Besides, to encourage the progressive evolution of intermediate semantic and boundary features $F_i^{s'}$ and F_i^b, we also attach additional prediction heads on them with the supervision from their individual ground truths.

3.2 Selective Mutual Evolution (SME) Module

We firstly introduce the key motivation of our reciprocal feature co-evolution mechanism. Compared with the segmentation of other daily objects, glass-like objects are more difficult to be distinguished from background regions, mainly due to their transparent property. One feasible workaround is trying to exploit useful boundary cues as assistance, which is also consistent with the human visual perception mechanism. In such a way, the semantic features around potential boundary will be enhanced and get more attention, which helps the model to capture the extent of glass region. However, the separation boundary between glass and surroundings is mostly too tiny to conduct an accurate prediction. Inspired by human's visual attention mechanism, semantic information of glass objects can be used to suppress false glass boundaries and highlight the features around real glass boundaries. Therefore, we propose to encourage feature co-evolution between boundary branch and semantic branch, simultaneously exploiting boundary cues to assist semantic features and exploiting semantic cues to assist boundary features. In a short word, a more accurate boundary prediction produces a better glass segmentation, which in turns facilitates a more accurate boundary map, and vice versa.

We then introduce the implementation details of our SME module. As shown in Figure 3, each basic mutual block takes F_{in}^s and F_{in}^b as inputs and outputs corresponding features F^s and F^b. Each block generates a two-channel attention map A from the joint feature representation of F_{in}^s and F_{in}^b, and mutually enhances the input features using the attention map to capture complementary information from each other.

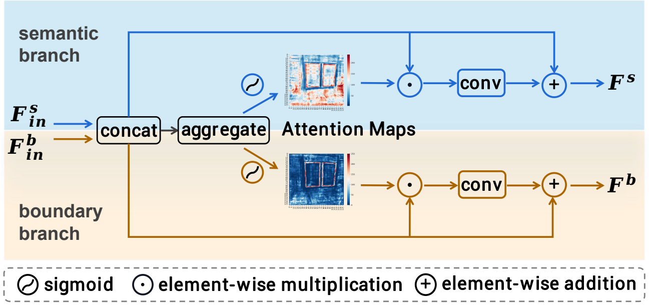

Figure 3: Illustration of the basic mutual block in our SME module. The attention maps are generated by a multi-branch aggregation operation on the semantic and boundary features, which helps to produce a more reliable attention score. Best viewed in color.

Specifically, the attention maps A are generated with a multi-branch aggregation operation on the concatenated features F_{in}^s and F_{in}^b. The concatenated feature is fed into two branches. We first use a convolution with kernel size of 3 to gather the local spatial context and then fuse the semantic and boundary information along the channel dimension to ensure a comprehensive view. Secondly, the fused feature is fed into two branches with convolutions of different kernel sizes (5 and 9). The large kernel based two-branch is designed to capture more long-range glass cues, which ensures a more reliable attention score. Finally, we stack several convolutions and a *sigmoid* operation on the concatenation of above two features to predict the attention maps A. The above process can be formulated as:

$$A = \left[\ a^s; a^b\ \right] = \sigma(Aggregate([F_{in}^s; F_{in}^b])), \tag{3}$$

where $A \in [0, 1]^{2\times h\times w}$, $Aggregate$ represents the aggregation operation, and $[\cdot]$ represents channel-wise feature concatenation. Then each channel of A (*i.e.*, a^s and a^b) is used as the attention map to enhance F_{in}^s and F_{in}^b respectively in a residual manner:

$$F^u = conv(F_{in}^u \bigodot a^u) + F_{in}^u, u \in \{s, b\}, \tag{4}$$

where $\bigodot$ denotes to element-wise multiplication. In this way, useful complementary information from either branch can be effectively exploited to adaptively enhance the features in the other branch without losing the original content.

We visualize the predicted two attention maps in Figure 3 to provide an intuitive explanation of our SME module. As shown in the figure, for the semantic branch, the features around the boundary are highlighted to accurately determine the distribution range of glass area. Notably, some regions outside the glass also receive relatively high attention. We assume that the network could attentively focus on some useful contextual information. As for the boundary branch, irrelevant contours in background regions are suppressed by relatively weak attention score to mitigate the noise disturbance, which is also in line with our analysis.

3.3 Structurally Attentive Refinement (SAR) Module

Despite the effectiveness of our SME module, there are still some difficult pixels remaining confusing to be distinguished, such as pixels located at boundaries, reflective regions and background regions with smooth surface. The inference of those ambiguous and difficult points is challenging without extra context priors. Fortunately, for any given difficult point, there are still several useful context cues to assist its inference, such as distance from its nearest boundary, curvature of

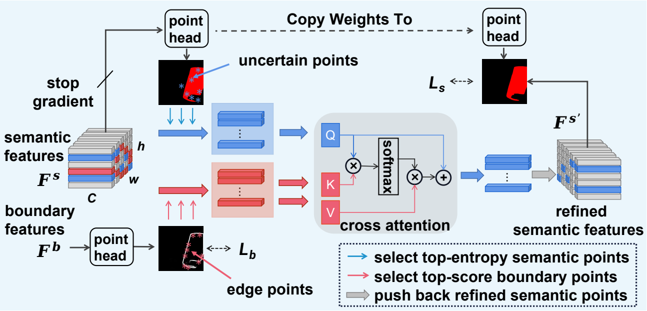

Figure 4: The illustration of our SAR module. The features of uncertain points are refined by global shape context.

local boundary and the whole shape of glass, *etc.*. Therefore, inspired by human's visual reasoning mechanism, we propose to adaptively aggregate the glass shape context as structural priors to help the inference of those ambiguous points. We firstly sample the features located at the top-confidence *boundary points* to construct a feature set as a compact representation of the glass shape context. Then we dynamically select the most *uncertain points* to conduct feature refinement along with the learning process, which acts as a kind of hard sample mining strategy. The refinement process is conditioned on the content around the current uncertain points, and is guided by the feature similarity with those boundary points. With such a refinement design, those uncertain points are allowed to freely explore useful shape context cues without the constraint from limited receptive field.

Following the above methodology, we propose a Structurally Attentive Refinement (SAR) module, as illustrated in Figure 4. The SAR module accepts the semantic feature F^s and boundary feature F^b from SME module as inputs, and outputs refined semantic feature $F^{s'}$. Based on the original input semantic features F^s, we can obtain initial glass prediction $P_s \in \mathbb{R}^{n\times h\times w}$ with n-class glass prediction head, and initial boundary map $P_b \in \mathbb{R}^{1\times h\times w}$ with edge prediction head. We use the prediction scores in P_s and P_b to select those uncertain points and top-confidence boundary points. Then we propose attentive feature refinement to enhance features of uncertain points under the assistance of top-confidence boundary points. After the features of those uncertain points are enhanced, we push them back into F^s according to their original indices to get the refined feature $F^{s'}$.

Attentive feature refinement. we introduce the details of the core attentive refinement operation as follows. 1) Firstly, we iterate over all pixels in P_s to calculate their Entropy for the measurement of prediction uncertainty. Then we select K points with the top entropy as the most ambiguous points set, the semantic features Q of which are later refined with extra shape context. 2) Secondly, we select M boundary points with the top prediction confidence based on P_b to construct the boundary feature set V. Intuitively, the feature set V could be considered as a compact representation of geometric information of glass, providing useful shape context. 3) Finally, for each feature in uncertain semantic feature set Q, we adaptively aggregate the most relevant shape context feature from boundary feature set V based on the feature correlation. We then obtain the refined features by fusing the aggregated results in a residual manner, which are later used for final semantic prediction. Notably, the above mentioned content-aware attentive refinement process can be easily implemented through the off-the-shelf cross-attention module, if we use the semantic features Q as queries and use boundary features V as both keys and values (*i.e.*, K=V). There are usually several parallel attention heads in one cross-attention module, and each of which could be formulated as following:

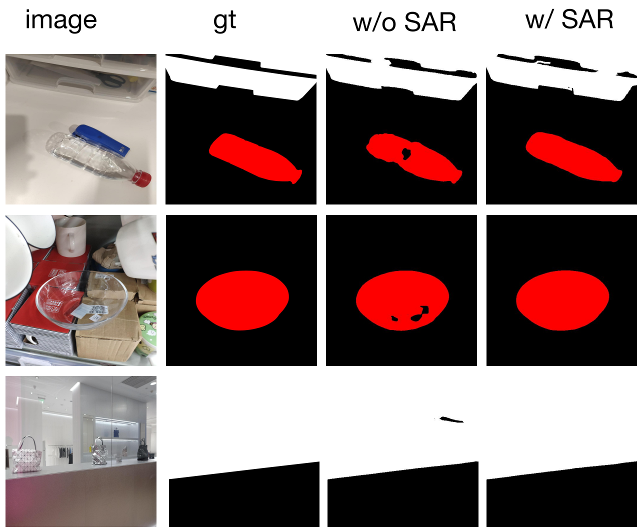

Figure 5: Visualized exhibition for the effectiveness of SAR module. The first and second columns are original images and their ground truths, and the third and fourth columns are the predicted maps without or with SAR module. Those ambiguous points, such as boundary or reflective points, can be refined with SAR module.

$$Attention(q,k,v) = softmax(\frac{q * k^t}{\sqrt{d_k}}) \cdot v, \tag{5}$$

where d_k is the dimension of input features q, k and v.

3.4 The Cascaded Connection

To fully explore the multi-scale feature representation within the backbone, we integrate our SME and SAR module into a cascaded structure. We denote the semantic and boundary features output from SME and SAR in stage i to F_i^s, $F_i^{s'}$ and F_i^b, respectively. As shown in Equation 1 and 2, we sequentially stack the two modules from higher level to lower level stage, making a cascaded optimization style. Besides, to aggregate more detailed information, F_1 from backbone is always concatenated with F_{i+1}^b before fed into SME module at each stage i. The semantic features $F_i^{s'}$ output from every stage are concatenated to produce the final semantic map. At the same time, the intermediate predictions of each stage are also supervised to encourage a progressive feature evolution.

3.5 Loss Design

Our RFENet is supervised with a joint function of semantic loss and boundary loss, which can be formulated as:

$$L = L_{s_{out}} + \lambda_s L_{s_i} + \lambda_b L_{b_i}, i \in \{1,2,3,4\}, \tag{6}$$

where the semantic losses $L_{s_{out}}$ and L_{s_i} are Cross-Entropy Losses and supervise the predictions from both the final glass prediction head, and the intermediate predictions of each stage in semantic branch. Meanwhile, the boundary losses L_{b_i} supervise the intermediate predictions of each stage in boundary branch. Considering that the boundary points only account for a small range in an image, we use the Dice Loss [Milletari *et al.*, 2016] as the boundary loss L_{b_i} to avoid the sampling imbalance problem. We generate the ground-truth boundary maps following [Xie *et al.*, 2020] with the thickness of 8. The λ_s and λ_b are used to balance the effect from L_s and L_b, which are set to 0.01 and 0.25 for all experiments.

4 Experiment

4.1 Datasets

Glass Datasets: (1) Trans10k [Xie *et al.*, 2020] is a large-scale transparent object segmentation dataset, consisting of 10,428 images with three categories: things, stuff and background. Images are divided into 5,000, 1,000 and 4,428 images for training, validation and test, respectively. It is by far the largest transparent object segmentation dataset with the most detailed annotations. Taking into consideration of the data amount and scene diversity, we conduct most of our experiments on this dataset to ensure a convincing result. **(2) GSD** [Lin *et al.*, 2021] is a medium-scale glass segmentation dataset containing 4,098 glass images, covering a diversity of indoor and outdoor scenes. All the data are randomly split into a training set with 3,285 images and a test set with 813 images. We use GSD to validate the generalization ability of our method.

Mirror Dataset: PMD [Lin *et al.*, 2020] is a large-scale mirror dataset contains 5,096 training images and 571 test images. It contains a variety of real-world images that cover diverse scenes and common objects, making it much closer to practical application. We conduct experiments on PMD dataset to demonstrate our model's transferability for mirror segmentation although it is designed for glass-like objects.

4.2 Evaluation Metrics

We follow the previous works to mainly adopt the following metrics to evaluate the performance of our model: mean Intersection over Union (mIoU), Pixel Accuracy (Acc), Mean Absolute Error (mAE), mean Balance Error Rate (mBER) and F-score. The mIoU is widely used to calculate the ratio of true positive prediction. The mBer measures a more comprehensive error rate by taking the sample imbalance problem into consideration. The Acc is used to provide a rough estimation of the pixel-level classification ability. And the mAE provides a measurement for the absolute prediction error of the segmentation map. Besides, we also follow [Yang *et al.*, 2019; Mei *et al.*, 2020] to measure F-score in PMD and GSD benchmark, which gives a more comprehensive view of Precision and Recall rate.

4.3 Implementation Details

We implement RFENet using the PyTorch framework [Paszke *et al.*, 2019]. The backbone network is initialized with ImageNet pre-trained weight, while the remaining parts are randomly initialized. ResNet50 is used for Trans10k and ResNeXt101 is used for GSD and PMD. We launch the training process on 4 GPUs with synchronized batch normalization, unless otherwise mentioned. For simplicity, we use stochastic gradient descent (SGD) as optimizer, which is scheduled with a poly policy with a power of 0.9.

Model	mIoU↑	Acc↑	mAE↓	mBER↓
BiSeNet [2018]	73.93	77.92	0.140	13.96
DenseAspp [2018]	78.11	81.22	0.114	12.19
DeeplabV3+ [2018b]	84.54	89.54	0.081	7.78
FCN [2015]	79.67	83.79	0.108	10.33
PSPNet [2017]	82.38	86.25	0.093	9.72
Translab [2020]	87.63	92.69	0.063	5.46
EBLNet(OS16) [2021]	89.58	93.95	0.052	4.60
EBLNet(OS8) [2021]	90.28	94.71	0.048	4.14
Ours(OS16)	**90.97**	**96.32**	**0.044**	**3.82**
Ours(OS8)	**91.25**	**96.50**	**0.043**	**3.68**

Table 1: Quantitative comparison between the proposed RFENet and state-of-the-art methods on Trans10k test set. OS means the output stride in the backbone network.

For Trans10k dataset, input images are resized to a size of 512×512 for both training and testing. The initial learning rate is set to 0.04, and weight decay is set to 0.0001. We use a mini-batch size of 4 for each GPU and run for 60 epochs.

For GSD dataset, following the same setting as [Lin *et al.*, 2021], the input images are firstly resized to 400×400 and then randomly cropped to 384×384. Random flipping is used for training. During inference, the test images are also first resized to 384 × 384 before fed into the network. The initial learning rate is set to 0.01, and weight decay is set to 0.0005. We run for 80 epochs with a batch size of 6 for each GPU.

For PMD dataset, we adopt the same setting as PMDNet [Lin *et al.*, 2020], where the input images are resized to 384×384. The initial learning rate is set to 0.03. The other settings remain the same as those on GSD dataset.

4.4 Comparison With the State-of-the-Arts

To demonstrate the superiority of our method, we conduct extensive experiments on three datasets, where we compare with recent state-of-the-art methods of glass-like object segmentation as well as some representative methods in common objects semantic segmentation. The semantic segmentation candidates to be compared are selected by referring to [He *et al.*, 2021].

Quantitative Evaluation. Firstly, as shown in Table 1, off-the-shelf methods that are designed for common objects produce inferior performance, such as the best one DeeplabV3+(ResNet50). This is in line with our analysis that the special transparent property of glass-like objects make it challenging to directly segment out without extra assistance.

Secondly, we make a comparison with the two recent strong competitors that are also designed for glass-like objects, *i.e.*, Translab [Xie *et al.*, 2020] and EBLNet [He *et al.*, 2021]. As shown from Table 1, our RFENet with the typical output stride of 16 achieves an impressive improvement with at least 1.5% gain in mIoU. Besides, when we adopt a finer feature map resolution of stride 8, our RFENet sets new

Model	mIoU↑	F_β↑	mAE↓	mBER↓
GDNet [2020]	79.01	0.869	0.069	7.72
Translab [2020]	74.05	0.837	0.088	11.35
GSD [2021]	83.64	0.903	0.055	**6.12**
PGSNet [2022]	83.65	0.868	0.054	6.25
Ours	**86.50**	**0.931**	**0.048**	6.23

Table 2: Quantitative comparison between the proposed RFENet and state-of-the-art methods on GSD test set.

Model	mIoU↑	F_β↑
MirrorNet [2019]	58.51	0.748
PMDNet [2020]	66.05	0.792
VCNet [2022]	68.25	0.812
Guan *et al.* [2022]	66.84	0.844
Ours	**73.56**	**0.851**

Table 3: Quantitative result of extending our RFENet to mirror dataset PMD and comparison with state-of-the-art mirror segmentation methods.

SME	SAR	Cascade	mIoU↑	Acc↑	mAE↓	mBER↓
			88.37	95.22	0.057	4.89
✓			89.70	95.56	0.050	4.33
✓		✓	90.07	95.86	0.048	4.15
	✓		88.66	95.27	0.056	4.69
✓	✓		90.04	96.07	0.049	4.02
✓	✓	✓	**90.42**	**96.10**	**0.046**	**3.99**

Table 4: The effects of SME and SRA module. We use a naive two-stream (*i.e.*, semantic and boundary) network as our baseline method.

semantic attention	boundary attention	mIoU↑	Acc↑	mAE↓	mBER↓
		88.37	95.22	0.057	4.89
✓		89.41	95.41	0.052	4.34
	✓	88.65	94.40	0.056	4.69
✓	✓	**89.70**	**95.56**	**0.050**	**4.33**

Table 5: The analysis of mutual learning mechanism. Both the two one-way assistance result in inferior performance.

records for all metrics on the Trans10k dataset. This promising improvement demonstrates the effectiveness of our core claim that encouraging the feature co-evolution between semantic branch and boundary branch helps to maximize the exploitation of their complementary information.

Thirdly, as shown in Table 2, on the GSD dataset, our RFENet achieves an improvement of 3.4% in terms of mIoU, compared with previous SOTA method GlassNet [Lin *et al.*, 2021]. The consistent improvement and competitive performance on GSD dataset demonstrates the generalization ability of our RFENet on other datasets, which is crucial for practical applications. We further extend our glass segmentation method to a mirror dataset, the PMD dataset. As shown in Table 3, our RFENet achieves even more improvement on mIoU, which clearly demonstrates the transferability of our method.

Qualitative Evaluation. Our method also achieves superior qualitative results which we exhibit in the appendix.

4.5 Ablation Study

In this section, we conduct extensive ablation study to demonstrate the effectiveness of SME and SAR modules. All experiments are conducted on a single GPU for efficiency.

Effectiveness of SME module. We use a two-stream network as our baseline method, in which we directly attach the glass prediction head and edge prediction head to the backbone feature F^s_{in} and F^b_{in}. As shown in Table 4, we firstly add SME$_4$ to conduct a single-scale mutual learning, which achieves a significant improvement with 1.5% in mIoU.

Notably, similar improvement from the SME module can also be achieved even if we have added the SAR module, which implies that the two proposed modules work in a complementary way. These quantitative results strongly demonstrate the effectiveness of the feature co-evolution between semantic feature and boundary feature.

For a more in-depth analysis on the mutual learning mechanism, we implement the one-way assistance by replacing the attentive feature enhancement operation in our SME module with an identity connection. To avoid any potential information flow from the other side, we also stop the gradient backpropagation from the generated attention map. As shown in Table 5, both the two one-way assistance strategies produce inferior performance, compared with the bi-directional assistance. It is worth noting that the combination of the two one-way assistance achieves much more improvement than the summation of their individual improvement, which shows the benefit of the feature co-evolution.

For further qualitative analysis of the attention map, we illustrate it in the appendix.

Effectiveness of SAR module. As shown in Table 4, the SAR module can achieve a further improvement in all metrics on the basis of the SME module. Similar improvement can also be consistently observed for the two-stream baseline model. Since the SME module only provides a local feature enhancement, these results effectively demonstrate the indispensability of assistance from global shape context. Qualitatively, from Figure 5 we can clearly see that without the SAR module to select and refine uncertain points, there are indeed some points that are difficult to predict correctly.

Effectiveness of cascaded connection. As shown in Table 4, the cascaded connection achieves consistent improvement. It implies the indispensability of multi-scale representation.

5 Conclusion

In this paper, we tackle the challenging problem of glass-like object segmentation with our proposed RFENet. The model contains two novel modules, Selective Mutual Evolution module for reciprocal feature learning between semantic and boundary branch, and Structurally Attentive Refinement module for refining those ambiguous difficult points through the global shape prior. Extensive experiments show that our model achieves state-of-the-art performance on Trans10k, GSD and PMD datasets.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61972157, No. 72192821), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Science and Technology Commission (21511101200), Shanghai Sailing Program (22YF1420300, 23YF1410500), CCF-Tencent Open Research Fund (RAGR20220121), Young Elite Scientists Sponsorship Program by CAST (2022QNRC001).

References

[Chen *et al.*, 2018a] Guanying Chen, Kai Han, and Kwan-Yee K Wong. Tom-net: Learning transparent object matting from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9233–9241, 2018.

[Chen *et al.*, 2018b] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[Ding *et al.*, 2019] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6819–6829, 2019.

[Guan *et al.*, 2022] Huankang Guan, Jiaying Lin, and Rynson WH Lau. Learning semantic associations for mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5941–5950, 2022.

[He *et al.*, 2010] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. In *European conference on computer vision*, pages 1–14. Springer, 2010.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2021] Hao He, Xiangtai Li, Guangliang Cheng, Jianping Shi, Yunhai Tong, Gaofeng Meng, Véronique Prinet, and LuBin Weng. Enhanced boundary learning for glass-like object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15859–15868, 2021.

[Huo *et al.*, 2022] Dong Huo, Jian Wang, Yiming Qian, and Yee-Hong Yang. Glass segmentation with rgb-thermal image pairs. *arXiv preprint arXiv:2204.05453*, 2022.

[Huynh *et al.*, 2021] Chuong Huynh, Anh Tuan Tran, Khoa Luu, and Minh Hoai. Progressive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16755–16764, 2021.

[Ji *et al.*, 2023] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, 20(1):92–108, 2023.

[Kirillov *et al.*, 2020] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020.

[Krähenbühl and Koltun, 2011] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011.

[Li *et al.*, 2018] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *Proceedings of the european conference on computer vision (ECCV)*, pages 355–370, 2018.

[Lin *et al.*, 2020] Jiaying Lin, Guodong Wang, and Rynson WH Lau. Progressive mirror detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3697–3705, 2020.

[Lin *et al.*, 2021] Jiaying Lin, Zebang He, and Rynson WH Lau. Rich context aggregation with reflection prior for glass surface detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13415–13424, 2021.

[Liu *et al.*, 2022] Jiang-Jiang Liu, Qibin Hou, Zhi-Ang Liu, and Ming-Ming Cheng. Poolnet+: Exploring the potential of pooling for salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):887–904, 2022.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[Mei *et al.*, 2020] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Don't hit me! glass detection in real-world scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3687–3696, 2020.

[Milletari *et al.*, 2016] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016.

[Pang *et al.*, 2020] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9413–9422, 2020.

[Paszke *et al.*, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca

Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[Qin *et al.*, 2019] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7479–7489, 2019.

[Takikawa *et al.*, 2019] Towaki Takikawa, David Acuna, Varun Jampani, and Sanja Fidler. Gated-scnn: Gated shape cnns for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5229–5238, 2019.

[Tan *et al.*, 2022] Xin Tan, Jiaying Lin, Ke Xu, Pan Chen, Lizhuang Ma, and Rynson WH Lau. Mirror detection with the visual chirality cue. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Wang *et al.*, 2019] Wenguan Wang, Shuyang Zhao, Jianbing Shen, Steven CH Hoi, and Ali Borji. Salient object detection with pyramid attention and salient edges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1448–1457, 2019.

[Xie *et al.*, 2020] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *European conference on computer vision*, pages 696–711. Springer, 2020.

[Xu *et al.*, 2015] Yichao Xu, Hajime Nagahara, Atsushi Shimada, and Rin-ichiro Taniguchi. Transcut: Transparent object segmentation from a light-field image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3442–3450, 2015.

[Yang *et al.*, 2018] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3684–3692, 2018.

[Yang *et al.*, 2019] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson WH Lau. Where is my mirror? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8809–8818, 2019.

[Yu *et al.*, 2018] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

[Yu *et al.*, 2022] Letian Yu, Haiyang Mei, Wen Dong, Ziqi Wei, Li Zhu, Yuxin Wang, and Xin Yang. Progressive glass segmentation. *IEEE Transactions on Image Processing*, 31:2920–2933, 2022.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[Zhao *et al.*, 2019] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8779–8788, 2019.

[Zheng *et al.*, 2022] Chengyu Zheng, Ding Shi, Xuefeng Yan, Dong Liang, Mingqiang Wei, Xin Yang, Yanwen Guo, and Haoran Xie. Glassnet: Label decoupling-based three-stream neural network for robust image glass detection. In *Computer Graphics Forum*, volume 41, pages 377–388. Wiley Online Library, 2022.

[Zhuge *et al.*, 2022] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.