Glass Segmentation with Multi Scales and Primary

Prediction Guiding

Zhiyu Xu[1] and Qingliang Chen[2]

[1] Jinan University, Guangzhou, China

Abstract. Glass-like objects can be seen everywhere in our daily life which are very hard for existing methods to segment them. The properties of transparencies pose great challenges of detecting them from the chaotic background and the vague separation boundaries further impede the acquisition of their exact contours. Moving machines which ignore glasses have great risks of crashing into transparent barriers or difficulties in analysing objects reflected in the mirror, thus it is of substantial significance to accurately locate glass-like objects and completely figure out their contours. In this paper, inspired by the scale integration strategy and the refinement method, we proposed a brand-new network, named as MGNet, which consists of a Fine-Rescaling and Merging module (FRM) to improve the ability to extract spatially relationship and a Primary Prediction Guiding module (PPG) to better mine the leftover semantics from the fused features. Moreover, we supervise the model with a novel loss function with the uncertainty-aware loss to produce high-confidence segmentation maps. Unlike the existing glass segmentation models that must be trained on different settings with respect to varied datasets, our model are trained under consistent settings and has achieved superior performance on three popular public datasets.

Keywords: glass segmentation

1   Introduction

Glass objects which possess two main characteristics, transparency and specular reflection, are known to be very hard for segmentation mainly due to two problems. The first problem is that these two attributes endow glass with substantial challenge in visual recognition, making it susceptible to blending with the background. The second problem is that the presence of indistinct and gradient boundaries further complicate the detection and segmentation process. In practices, failure to detect glasses can potentially lead to collisions with their barriers for machines such as robots and autonomous vehicles, as well as hindering the tracking of objects. Consequently, new powerful methods and models for the detection and segmentation of glasses are badly needed.

The existing segmentation methods focus on either transparent objects or mirror objects. For transparent materials, existing methods mainly leverage boundary detection for assistance ([6], [26]). Furthermore, some methods have been proposed to detect and

segment glass using new attributes such as spectral polarization [25]. However, it is evident that these methods introduce additional computational overhead, and these attributes are not universally present in every glass object. For instance, some mirrors have well-defined boundaries, while certain glass exhibits perplexing edges. Additionally, acquiring other attributes, such as spectral polarization information, for each object is not always convenient. Actually, transparent objects and mirrors can be classified as the same substance, glass. Therefore, combining these two tasks intuitively makes a lot of sense and can significantly improve the applicability of the model.
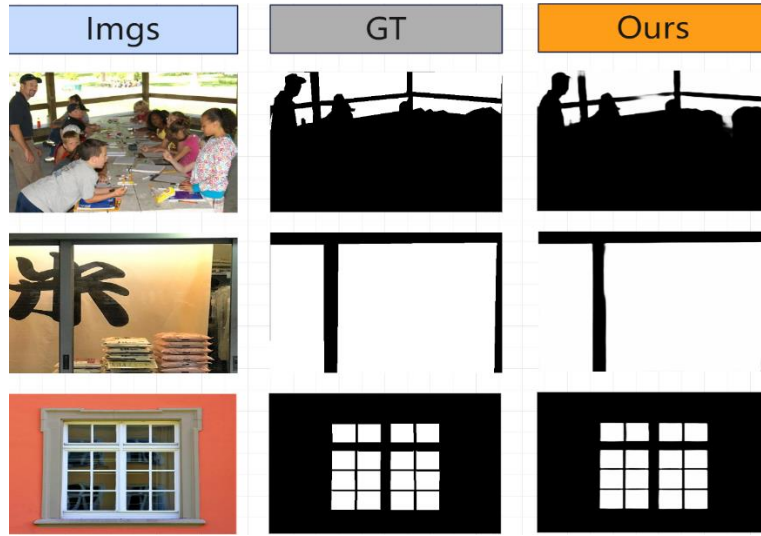

Fig. 1. Visualization of our predicted results. Compared with the ground truths, our segmentation maps exhibit fine, complete, and strong discriminative properties, which is mainly contributed by our combination of multi-scales and refinement strategies. The predictions of those ambiguous regions also gain the help from UAL.

Considering these two restriction, we intend to combine transparent objects segmentation and mirror segmentation into glass segmentation and rely solely on conventional RGB information, similar to the approach taken for other objects. This approach not only simplifies the data acquisition and processing but also facilitates a better analysis, thereby improving model applicability.

In this paper, we are dedicated to addressing the problems arising from object interference within glasses and the challenging segmentation of objects due to the special boundaries of glasses in the background. Drawing inspiration from multi-scale input strategies and iterative refinement modules, we propose a novel network architecture, referred to as MGNet, which includes a Fine-Rescaling and Merging module (FRM) and a Primary Prediction Guiding module (PPG). FRM makes slight adjustments to the original input image's dimensions, enhancing the spatial relationship among objects

within the image and capturing more abstract representations. As the two problems mentioned above, the presence of reflected objects in mirrors and items located behind transparent glass introduces the likelihood of two distinct groups within an image – one representing reality and the other created by the glass. The application of scale-space theory, which seeks to comprehensively address natural variations in scales to optimize the understanding of image structure offering a promising avenue for discerning the spatial relationship between the real world and the glass-induced reflections.

Furthermore, this strategy significantly enhances model performance by effectively handling vague and confusing edges, as outlined in three versions tailored to address this challenging detection scenario. And PPG aims to address the challenge caused by the second problem of incomplete and low-confidence segmentation posed by the ambiguous boundaries between the glass-like objects and the background, mining the leftover semantics from the final features by guiding the segmentation process with the primary prediction. Furthermore, to enhance the effectiveness of eliminating the uncertainty regions (low-confidence regions), we utilize the Uncertainty-Aware Loss (UAL) proposed by [28] as an auxiliary loss function.

Our contributions can be summarized as follows: (i) We propose MGNet, a novel glass-like object segmentation model which combines two empirically effective strategies of multi-scales and refinement with a rather strong purpose to address the two problems mentioned above. (ii) We propose a Fine-Rescaling and Merging module (FRM) to improve the ability of our model to extract spatially relations and a Primary Prediction Guiding module (PPG) to better mining the leftover semantics from the fused features. (iii) Our methods combines transparent objects and mirror segmentation and has been implemented, achieving state-of-the-art performance on two popular benchmarks (two different tasks) and superior performance on another instance transparent objects segmentation dataset under the consistent train settings, presenting better transferability since existing models must adjust their settings respect to each benchmark for only one task. (iv) our model solely rely on RGB-photos and beats those models with special physical property, showing a great possibility of utilizing RGB-photos to excellently segment transparent objects and mirror.

2 Related Works

Glass-like Object Segmentation. Segmenting objects with a glass-like appearance presents a significantly greater challenge compared to commonly seen objects and this heightened difficulty arises primarily from the fact that the inner regions of glass objects often exhibit a perplexing similarity to their surrounding backgrounds. To address this issue, some methods [25] have turned to the utilization of additional multi-modal information, such as 4D light-field data, refractive flow maps, thermal imaging and spectral polarization. Regrettably, the acquisition of such multi-modal data is relatively expensive, thus limiting its broader applicability. Instead, recent contributions from researchers such as ([16], [17], [26], [37]), have led to the creation of large-scale RGB image datasets specifically tailored to glass-like objects, promoting this research in the community. However, due to the unique characteristics of glass-like objects,

conventional existing semantic segmentation methods ([3], [46]), have failed to deliver promising results. Similarly, many cutting-edge approaches in salient object detection ([28], [29]) have also struggled, as glass objects may not always exhibit salient features. Additionally, ([16], [37]) have introduced methods for the segmentation of glass-like objects with the aid of boundary cues, leveraging the precise localization afforded by boundaries. Unlike the methods mentioned above, our approach only processes the raw RGB images with a novel network combining two strategies of multi-scales and refinement without any boundary computation.

Scale Space Integration. Scale-space theory typically aims to address natural variations in scales by achieving an optimal comprehension of image structure, providing an exceedingly effective and theoretically sound framework. This concept has found widespread applications in computer vision, including the utilization of image pyramids [1] and feature pyramids ~\cite{lin2017feature}. Recent CNN-based methods for Camouflaged Object Detection (COD) ([5], [28]) and Salient Object Detection (SOD) ([12], [27], [29]) have explored strategies that combine inter-layer features to enhance feature representations. These approaches have demonstrated positive impacts on accurate object localization and segmentation. However, for the glass segmentation task, the existing approaches overlook the substantial effect of varying input scales on detecting glass-like objects and thus they are often confused by the complex spatial relationships in images between objects. In this paper, we apply the mixed-scale integration strategy to adjust the attention of our model to the scales it needs most with respect to each pixel to significantly improve the ability of our model to figure out the structure of input images. Furthermore, we design our network along with the fine-rescale and merging module and the hierarchical channel-down decoder which effectively extract and fuse the semantics from different scales.

Iterative Refinement. The iterative refinement aims to optimize the secondary structure and elaborate details through repeating the final or nearly final process of feature extraction supervised by the results predicted in the last step. Recently, this strategy has been widely utilized on few-shot learning [44] and camouflaged object detection as well [13]. Jia et al. propose an attention-based iterative refinement module in [13], which samples the target area iteratively to improve the ability of the model to detect the objects in small scales. Unlike the models mentioned above, our proposed iterative refinement module intends to make our model segment objects in a coarse way firstly and then guide the primary predicted results with the final features to fully mine the semantics left in the processed final features. Therefore our model is able to detect the severely glass-like objects and generates final prediction which is more structured and regular.

3 Methodology

In this section, we first elaborate on the overall architecture of the proposed MGNet, and then present the details of each module.

3.1 Overall Architecture

The overall architecture of the proposed MGNet is illustrated in Fig.2. Inspired by the rescaling strategy and the rechecking strategy of human beings, we better and advance those strategies for getting performance promotion in glass-like object segmentation. By consolidating information from various scales, we can effectively uncover
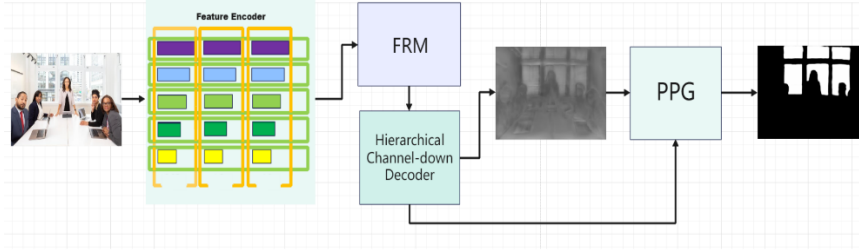

Fig. 2. The overall framework of MGNet. The fine-rescaling and merging module (FRM) is adopted to integrate full-channel features of different levels extracted by a shared encoder to mine the critical clues from different scales. The hierarchical channel-down decoder (HCDD) further enhances the feature discrimination by constructing a multi-path structure inside the features while reducing the channel of features. Then, a coarse probability map is generated and sent to the primary prediction guiding module (PPG) to complete the coarse-to-fine process for mining leftover semantics and uncertainty elimination. After refinement, a final probability map of the camouflaged object on the input image can be obtained. The times of guiding can be adjusted according to practical needs.

subtle but valuable cues in complex situations, ultimately aiding in the process of glass-like objects segmentation, thus we finely adjust the scales of input image (height and weight), i.e., 0.7x, 1.0x, 1.2x, which conform to our ideas of multi-scales input. The images of different scales share a feature encoder to extract different levels of features and the feature encoder we use in this paper is the pre-trained ResNeXt101. The features of different scales are organized based on their respective channels, and these groups are then directed to the Fine-Rescaling and Merging module (FRM), which is based on the attention-aware filtering mechanism to integrate features that contain rich scale-specific information. The module will generate attention to different scales which fits the picture being detected and significantly improve the model's ability to capture crucial and informative semantic cues for detecting challenging, glass-like objects. After that, the fused multi-level features are gradually integrated through Hierarchical Channel-Down decoder (HCDD) in a top-down manner to enhance the mixed-scale feature representation while their channel is reduced accordingly to be able to mix with others hierarchically. It further expands the receptive field ranges and integrates information from the semantic level to the local level. At the end of the integration process, a coarse logits map can be obtained. To eliminate the uncertainty in the coarse logits map and further mine the leftover information of structures and local details of the final fused features, we employ the primary prediction guiding module to process the feature generated by the last module (HCDD) with the primary logits map. After the refinement, the final segmentation map is generated by the subsequent logits map through a convolutional layer and it is supervised by the ground truth. The loss function is

composed of the binary cross entropy (BCE) loss and an uncertainty-aware loss (UAL) to enable the model to eliminate the uncertain regions and produce a structured, regular and accurate prediction.

3.2 Feature Encoder

We first extract deep features through a shared encoder and group them by their channels. For fully implementing our idea of finely adjusting scales to promote spatially relationship input,

the three scales are intuitively set to 0.7x, 1.0x, 1.2x of the height and weight. The feature encoder is constituted by the commonly-used ResNeXt101 where the structure after "layer4" is removed. For more details, there will be five groups of features after the feature encoder, whose channels are 2048, 1024, 512, 256 and 64, respectively and whose heights and width are 1/32, 1/16, 1/8, 1/4, 1/2 of the original input size, respectively. Next, these groups of features are fed successively to the fine-rescaling and merging module (FRM) and the hierarchical channel-down decoder (HCDD) for subsequent processing.
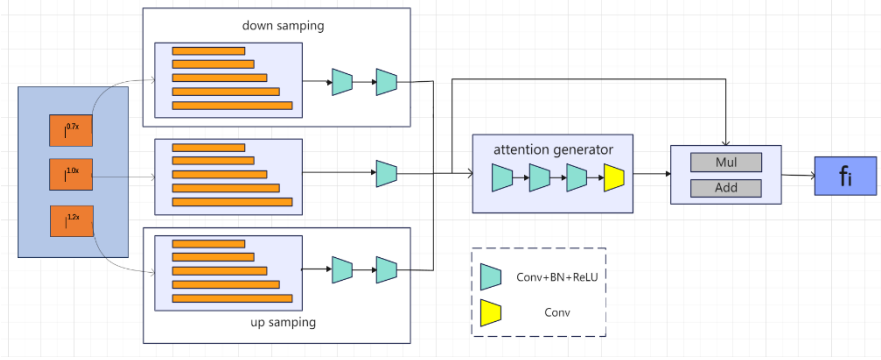

Fig. 3. Illustration of the full scale merging module (FRM).

3.3 Fine-Rescaling and Merging Module (FRM)

While nearly every object in a given image maintains a consistent spatial relationship with other objects, which remains unchanged when the image is rescaled, integrating features from different scales can greatly highlight this relationship and improve the scale robustness of our model to help address the first problem mentioned above. For the second problem, the identification of edges, crucial for detecting concealed objects, can be significantly affected by rescaling, particularly in the case of low resolution and high levels of image noise in the test pictures. This impact becomes particularly evident when dealing with confusing glass-like objects and those with indistinct edges. To address this problem, we introduce the Fine-Rescaling and Merging module (FRM).

The attention-based FRM aims to weight and combine scale-specific information, as shown in Fig.3. Following [28], the scale merging module is composed of several units and layers which implement filtering and aggregation to self-adaptively highlight the

expression of different scales. The original scales (1.0x) is resized to 0.7 and 1,2 times of the lengths to generate relatively even input scales. After that, there are three groups of features belonging to 0.7x, 1.0x, and 1.2x, respectively. For $f_i^{1.2}$, we further extract features by two stacked "Conv-BN-ReLU" layers with different receptive fields and then use a mixed addition of ``max-pooling'' and ``average-pooling'' to down-sample it, which helps to conserve the meaningful and various responses for glass-like objects in high-resolution features. For $f_i^{0.7}$, we further extract features by two stacked "Conv-BN-ReLU" layers with the same receptive fields and up-sample it by bi-linear interpolation. For $f_i^{1.0}$, it goes through a "Conv-BN-ReLU" layers. And these features are then fed into an attention generator stacked by three "Conv-BN-ReLU" layers and one convolutional layer.

A three-channel feature map is calculated through these layers and after a softmax activation layer, the attention map A^k ($k \in \{0.7, 1.0, 1.2\}$) corresponding to each scale can be obtained and used as respective weights for the final integration. The process is formulated as:

$$f_i^{0.7} = \zeta_1(f_i^{0.7}), f_i^{1.0} = \zeta_2(f_i^{1.0}), f_i^{1.2} = \zeta_3(f_i^{1.2}), \quad (1)$$

$$A_i = softmax(\xi([f_i^{1.2}, f_i^{1.0}, f_i^{0.7}], \theta)) \quad (2)$$

$$f_i = A_i^{1.2} * f_i^{1.2} + A_i^{1.0} * f_i^{1.0} + A_i^{0.7} * f_i^{0.7} \quad (3)$$

Where ζ_1 indicates the stacked "Conv-BN-ReLU" layers and the down-sample structure, ζ_2 is the stacked "Conv-BN-ReLU" layers, ζ_3 represents the stacked "Conv-BN-ReLU" layers and bi-linear interpolation, ξ denotes the stacked "Conv-BN-ReLU" layers in the attention generator, θ is the parameters of these layers, and [] indicates input features are processed after the concatenation operation.
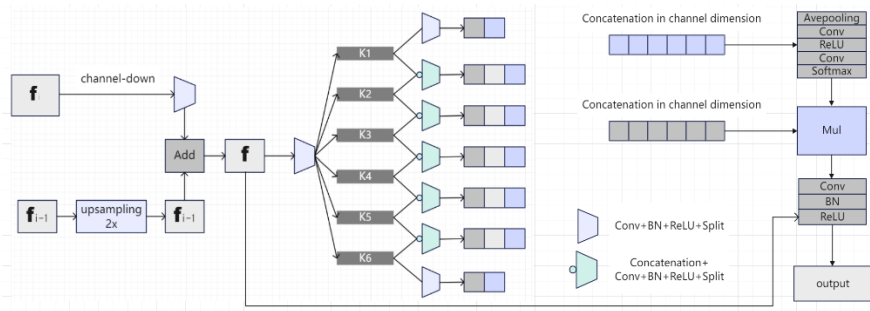

Fig. 4. Illustration of the hierarchical channel-down unit (HCDU).

3.4 Hierarchical Channel-Down Decoder (HCDD)

As multi-scales contain consistent compositions and accordant structures, after \textbf{FRM} integrates the multi-scale features into the original scale, the spatial and semantic relationships are highlighted in the feature representations. Similar to most cases, different channels also contain differentiated semantics. Thus, it is essential to mine valuable clues contained in different channels. Motivated by [28], we introduce Hierarchical Channel-Down Units (**HCDUs**) to interact and reorganise the critical information in different channels in a pyramid way, which strengthen features from coarse-grained group-wise iteration to fine-grained channel-wise modulation in the decoder, as shown in Fig.4. The input f_i of HCDU$_i$ is the fused feature f_{6-i} from FRM and the input f_{i-1} is the output of HCDU$_{i-1}$. We first reduce the channel of f_i by a stacked 3x3 "Conv-BN-ReLU" layer, up-sample f_i by the direct bi-linear interpolation and add the two features as f:

$$f = \delta(f_i) + f_{i-1}^{2.0X} \qquad (4)$$

where δ represents the stacked "Conv-BN-ReLU" layer.

To succinctly describe our unit, we define the stacked "Concatenation-Conv-BN-ReLU-Split" layers as CCBRS. The CCBRS processing the concatenation of the input two features is composed of a 3x3 convolutional layer, a batch normalization layer, a ReLU layer and a split layer breaking the feature into three or two chunks along the channel dimension. First, we expand the channels of f six times by a 1x1 "Conv-BN-ReLU" layer and split f into six chunks along the channel dimension (k1,k2,k3,k4,k5,k6) shown in Fig.4. Then the six features are successively processed in pair-wise groups through **CCBRS** and the unit processes the first and the last feature additionally by **CCBRS** without concatenation. The feature interaction is like a pyramid, as depicted in Fig.4. Every group will generate two or three sub-features, the unit concatenates the last sub-feature in every sub-features group as the weighted feature, denoted by tensor M. The feature then goes through an average-pooling layer, a ReLU layer between two 1*1 convolutional layers and an activation layer to generate the weight feature map. Afterwards, the unit concatenates the first sub-feature in every sub-feature group, denoted by tensor T and multiple the tensor T by the weight feature map. The final output feature can be obtained after a 3*3 convolutional layer, a batch normalization layer and a ReLU layer. The whole process can be formulated as:

$$\tilde{f} = \mathcal{R}eLU((f + \mathcal{N}(\mathcal{C}(\Psi(M) \cdot T)))) \qquad (5)$$

where ReLU, N, C represent the activation layer, the normalization layer and the convolutional layer, respectively and Ψ denotes the weight generator layer.

The Hierarchical Channel-down Decoder (HCDD) contains five HCDUs. A single-channel logits map can be obtained after we integrate the five fused features with HCDUs and process the integrated features by a 1x1 convolutional layer. A coarse confidence map P that somewhat highlights the glass-like objects can be then generated by a sigmoid function, as shown in Fig.2.
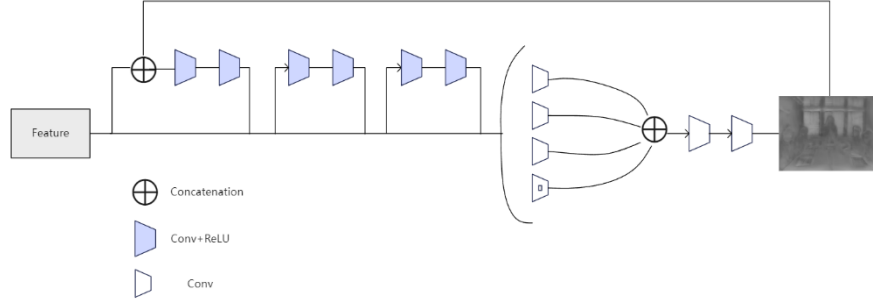
Fig. 5. Illustration of the primary prediction guiding module (PPG).

3.5 Primary Prediction Guiding Module (PPG)

As it is obvious that the initial prediction is an important reference and contains valuable clues about the rough position of the objects, we want our model to be able to guide further predictions with the original results. Thus, we introduce the primary prediction guiding module (PPG) to refine our prediction for detecting the severely confusing glass (mirror or glass baffle) and eliminating the uncertain regions in our segmentation map mainly due to vague boundaries, which is similar to the behaviour of human beings of checking more to confirm the structure and the edge to help address the second problem.

Specifically, as shown in Fig.5. The input feature of the iterative refinement module is the output feature of the HCDD. The final generated logits map can be represented as:

$$\mathcal{M}_t = \mathcal{F}(x, \mathcal{M}_{t-1}) \tag{6}$$

M represents the output of the module, t is the number of times and x is the output of the HCDD, which is unchanged during the refinement process. The function $\mathcal{F}$ can be described as follow : in the first residual block, the feature x concatenates with the logits map predicted in the last refinement process, and goes through the two 3x3 convolutional and ReLU layers (the first one has 33 filters and the second one has 32 filters) and is added with the original input feature. Output feature is denoted by **k1**. For the other two residual blocks, they are composed of two 3x3 convolutional and ReLU layers with 32 filters (the processed **k1** is denoted by **k2**) and **k2** is added with the original input feature **k1** (output feature is denoted by **k3**). After that, we apply the Atrous Spatial Pyramid Pooling module (ASPP) proposed in Deeplab V3 [2] to capture multi-scale information containing in **k3**. The module ASPP consists of four parallel branches that include three 3×3 convolutions with atrous rates of 6, 12, and 18, respectively and a 1×1 convolution. The 1×1 convolution is operated on the image-level feature which is achieved by global average pooling. The output features from 4 branches are concatenated and fused by another 1×1 convolution with 128 filters. The final single-channel logits map can be obtained after another 1x1 convolutional layer with 32 filters. This

primary prediction guiding module implements a rather coarse-grained to fine-grained process, to preserve the critical semantic clues and further detect the targeted regions.

3.6   Loss Functions

The binary cross entropy loss (BCE) which is widely used in image segmentation tasks is adopted as a part of our loss function and its mathematical form is:

$$l_{BCE}^{i,j} = -g_{i,j}logp_{i,j} - (1 - g_{i,j})log(1 - p_{i,j})$$

where $g_{i,j} \in \{0, 1\}$ and $p_{i,j} \in [0, 1]$ denote the ground truth and the predicted value at position (i,j), respectively.

To eliminate the uncertainties of the segmentation map in the iterative refinement module and generate high-confidence results, we use an auxiliary loss function of the BCE, i.e., the uncertainty-aware loss (UAL) proposed in [28]. In the final probability map of the camouflaged object, the pixel value range is [0,1], where 0 means the pixel belongs to the background, and 1 means it belongs to the camouflaged object. Therefore, the closer the predicted value is to 0.5, the more uncertain the determination about the property of the pixel is. The UAL maximizes at x = 0.5 and minimizes at x = 0 or x = 1 to optimize the ambiguity of the prediction. the UAL is formulated as

$$L_{UAL} = 1 - |2p_{i,j} - 1|^2$$

So the final loss function is as follows:

$$L = L_{BCE} + 1.5 * \lambda * L_{UAL} \qquad (7)$$

where λ ranged from zero to one is the balance coefficient which utilizes an increasing cosine strategy.

4   Experimental Results

4.1   Experimental Settings

Datasets. We evaluate the proposed method on three widely used glass and mirror datasets i.e., Trans10k[37], GSD [16] and PMD [17]. Trans10k 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. GSD 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. PMD is another large-scale mirror dataset containing 5,096 training images and 571 test images. It has a variety of real-world images that cover diverse scenes and common objects, making it much closer to practical applications.

Unlike existing methods, we trained our model with respect to each dataset under consistent settings, while the existing glass segmentation models must be trained on different settings with respect to varied datasets.

Evaluation Criteria. We apply the following evaluation metrics to assess the performance of our model: mean Intersection over Union (mIoU), Mean Absolute Error (mAE), and mean Balance Error Rate (mBER). The mIoU is commonly employed to determine the ratio of true positive predictions. The mBER offers a more inclusive assessment of error rates, accounting for sample imbalance and mAE measures the absolute prediction error of the segmentation map.

Implementation Details. The proposed MGNet is implemented with PyTorch. The encoder is initialized with the parameters of pre-trained ResNeXt101 and the remaining parts are randomly initialized. And SGD with momentum 0.9 and weight decay 0.0005 is chosen as the optimizer. The learning rate is initialized to 0.08 and follows a linear warmup and linear decay strategy. The entire model is trained for 32 epochs with a batch size of 12 in an end-to-end manner on an NVIDIA Tesla V100-SXM2-32GB GPU. During training and inference, the input scale is 384×384. Random flipping and rotating are employed to augment the training data.

Table 1. Quantitative comparison with 11 SOTA methods on GSD benchmark dataset.

Methods	Publication Venues	GSD (glass surface dataset)		
		IoU↑	MAE↓	BER↓
PSPNet [46]	CVPR2017	70.26	0.110	10.66
BDRAR [52]	ECCV2018	75.91	0.081	8.61
BASNet [30]	CVPR2019	69.79	0.106	13.54
MINet [29]	CVPR2020	77.29	0.077	9.54
SINet [5]	CVPR2020	77.04	0.077	9.25
GDNet [26]	CVPR2020	79.01	0.069	7.72
TransLab [37]	ECCV2020	74.05	0.088	11.35
GSDNet [16]	CVPR2021	83.64	0.055	6.12
PGSNet [25]	CVPR2022	83.65	0.054	6.25
SAM [14]	ICCV2023	50.60	0.213	23.91
MGNet(Ours)	—	84.08	0.051	6.32

Table 2. Quantitative comparison with 8 SOTA methods on PMD benchmark dataset.

Methods	Publication Venues	PMD (mirror dataset)	
		IoU↑	MAE↓
EGNet [47]	ICCV2019	60.17	0.036
BDRAR [52]	ECCV2018	58.43	0.043
PSPNet [46]	CVPR2017	60.44	0.039
DeepLabv3+ [3]	ECCV2018	64.08	0.040
PMDNet [17]	CVPR2020	62.40	0.087
MirrorNet [42]	ICCV2019	62.50	0.041
Guan et al. [8]	CVPR2022	66.84	0.049
SAM [14]	ICCV2023	64.75	0.052
MGNet	—	68.77	0.028

Table 3. Quantitative test of our model MGNet on Trans10k benchmark dataset.

Methods	Settings	Trans10K		
		IoU↑	MAE↓	BER↓
MGNet	**B+H+F+P+U**	89.97	0.030	4.05

4.2 Comparisons with State-of-the-art

To demonstrate the effectiveness of the proposed method, we compare it with several state-of-the-art glass-like object segmentation models, including mirror detection and transparent object detection. All the predictions of baselines are either provided by the authors or generated by open source models retrained.

Quantitative Comparison. Tab.1 summarizes the quantitative results of our proposed method against 10 baselines on the challenging glass surface benchmark dataset under three evaluation metrics. Tab.2 presents the quantitative results of our proposed method against 8 counterparts on the demanding mirror benchmark dataset under two evaluation metrics. We can see that our model has outperformed all the state-of-the-art baselines both in IoU and MAE and has comparable performance in BER.

Qualitative Evaluation. Our method also achieves excellent qualitative results which are exhibited in the Figure 1.

Table 4. Quantitative comparison between the complete model (MGNet) and ablative models on two benchmark datasets（GSD and PMD). And **B**, **F**, **H**, **P** and **U** denote backbone, FRM, HCDD, PPG and UAL, respectively. "+" means addition of the module. The results of our complete model (MGNet) are **bolded**.

Settings	GSD			PMD	
	IoU↑	MAE↓	BER↓	IoU↑	MAE↓
B+H+P	75.74	0.082	9.04	60.95	0.042
B+H+F	77.13	0.077	8.46	61.89	0.040
B+H+F+P	80.12	0.059	7.23	64.22	0.037
B+H+F+P+U (our setting)	**84.08**	**0.051**	**6.32**	**68.77**	**0.028**

4.3 Ablation Studies

To validate the effectiveness of the proposed modules in our model, we have also performed the following ablation studies on these benchmarks.

Effectiveness of FRM. Integrating multiple scales of input pictures has been found useful for understanding the complex spatial relationship of the glass-like objects with other objects. The FRM also enables our model to adjust attention to different scales with respect to each pixel. To confirm its effectiveness, we delete the module from MGNet, preserving other modules without any modification, train it under the same

settings and test it on two datasets (GSD and PMD), with the results shown in Table 4. It can be easily seen that the model has made great progress on all the metrics after incorporating the FRM to integrate the critical semantics of different input scales.

Effectiveness of PPG. Unlike existing iterative refinement strategies, our refinement strategy aims to implement a coarse-to-fine segmentation process and capture the structure and edge of the glass-like objects with the primary prediction, which is also the process of eliminating the uncertainty in the segmentation map. Our model first generates a coarse logits map which slightly highlights the glass-like objects but still preserves the most critical semantic clues and leaves many uncertain regions mainly due to the vague boundaries. To confirm its effectiveness of improving the performance of our model, we uninstall the primary prediction guiding module (PPG) from MGNet. It can be seen in Table 4 that with PPG, the model has obvious performance boost of the weighted mIoU, MAE and BER on the two datasets (GSD and PMD). The results demonstrate convincingly that our model successfully implement the idea of refinement with the primary prediction to get more accurate and complete answers. Then, based on this coarse logits map, the segmentation map are refined and becomes more and more regionally distinct.

Table 5. Quantitative comparison of the performance between models with different refining times.

Settings	GSD		PMD	
	IoU↑	MAE↓	IoU↓	MAE↓
MGNet-iter0	35.74	0.121	20.95	0.142
MGNet-iter1	**84.08**	**0.051**	**68.77**	**0.028**
MGNet-iter2	85.22	0.049	71.86	0.027
MGNet-iter3	86.12	0.048	73.17	0.026

Furthermore, Table 5 shows the change of the performance of our model on the GSD and PMD dataset with the variation of the refinement times from 0 to 3. And it can be observed that although our model is trained on the setting of once refinement, the test performance still improves when the refinement runs increase. The results demonstrate convincingly that our model successfully implements our idea of guiding to get more accurate and complete predictions.

Uncertainty-aware Loss. In the final probability map of the glass-like objects, the pixel value range is [0,1], where 0 means the pixel belongs to the background, and 1 represents it belongs to the camouflaged object. The closer the predicted value is to 0.5, the more uncertain the determination about the property of the pixel is. Also with the purpose of enabling our model to eliminate the uncertainty during refining, the UAL should maximize at x = 0.5 and minimize at x = 0 or x = 1 to optimize the ambiguity of the prediction. To confirm its effectiveness, we delete the UAL from our model, and train with the same settings. The results in Table 4 indicate that the performance of our model is generally enhanced after adding the UAL to the loss function, justifying its

efficacy. Furthermore, for another transparent object instances dataset Trans10k, Table 3 demonstrates the quantitative results of our proposed MGNet with all modules installed, showing that our model can achieve good performance on this dataset. Different from model comparison, the evaluation on this dataset is to test the model's capability to detect the presence of glass-like objects. Specifically, all ground truth annotations are binary, with areas containing glass-like objects marked in white and the background marked in black.

5   Conclusion

In this paper, we address the challenging problem of glass-like object segmentation with the proposed MGNet, which contains three novel modules of the fine-rescaling and merging (**FRM**) for integrating multi-scales features, the hierarchical channel-down decoder (**HCDD**) for further mining valuable clues contained in different channels, and the primary prediction guiding module (**PPG**) for refining more overall structures and detailed information. Extensive experiments show that our model can achieve superior performance on Trans10k, GSD and PMD datasets with respect to state-of-the-art counterparts. And the future work will explore more advanced techniques for extracting more nuanced and richer features for better segmentation.

References

1. Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. Pyramid methods in image processing. RCA engineer, 29(6):33–41, 1984.
2. Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Mur- phy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2017.
3. 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.
4. Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In Proceedings of the European conference on computer vision (ECCV), pages 234–250, 2018.
5. Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2777–2787, 2020.
6. Ke Fan, Changan Wang, Yabiao Wang, Chengjie Wang, Ran Yi, and Lizhuang Ma. Rfenet: Towards reciprocal feature evolution for glass segmentation. arXiv preprint arXiv:2307.06099, 2023.
7. Shang-Hua Gao, Yong-Qiang Tan, Ming-Ming Cheng, Chengze Lu, Yunpeng Chen, and Shuicheng Yan. Highly efficient salient object detection with 100k parameters. In European Conference on Computer Vision, pages 702–721. Springer, 2020.
8. 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.

9. Dongsheng Han, Chaoning Zhang, Yu Qiao, Maryam Qamar, Yuna Jung, SeungKyu Lee, Sung-Ho Bae, and Choong Seon Hong. Segment anything model (sam) meets glass: Mirror and transparent objects cannot be easily detected. arXiv preprint arXiv:2305.00278, 2023.

10. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyra-mid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9):1904–1916, 2015.

11. 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 recog-nition, pages 770–778, 2016.

12. Wei Ji, Ge Yan, Jingjing Li, Yongri Piao, Shunyu Yao, Miao Zhang, Li Cheng, and Huchuan Lu. Dmra: Depth-induced multi-scale recurrent attention network for rgb-d saliency detec-tion. IEEE Transactions on Image Processing, 31:2321–2336, 2022.

13. Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4713–4722, 2022.

14. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gus-tafson, Tete Xiao, Spencer Whitehead, Alexan-der C Berg, Wan-Yen Lo, et al. Segment anything. arXiv preprint arXiv:2304.02643, 2023.

15. Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In Proceed-ings of the IEEE conference on computer vision and pattern recognition, pages 5455–5463, 2015.

16. 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 Vi-sion and Pattern Recognition, pages 13415–13424, 2021.

17. Jiaying Lin, Guodong Wang, and Rynson WH Lau. Progressive mirror detection. In Pro-ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3697–3705, 2020.

18. Tsung-Yi Lin, Piotr Doll´ar, Ross Girshick, Kaiming He, Bharath Hariha-ran, and Serge Belongie. Feature pyramid networks for object detection. In Proceedings of the IEEE con-ference on computer vision and pattern recognition, pages 2117–2125, 2017.

19. Tony Lindeberg. Feature detection with automatic scale selection. International journal of computer vision, 30:79–116, 1998.

20. Tony Lindeberg. Scale-space theory in computer vision, volume 256. Springer Science & Business Media, 2013.

21. Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual at-tention for saliency detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3089–3098, 2018.

22. Yun Liu, Xin-Yu Zhang, Jia-Wang Bian, Le Zhang, and Ming-Ming Cheng. Samnet: Stere-oscopically attentive multi-scale network for lightweight salient object detection. IEEE Transactions on Image Processing, 30:3804–3814, 2021.

23. Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for se-mantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3431–3440, 2015.

24. Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In Proceedings of the IEEE Conference on computer vision and pattern recognition, pages 6609–6617, 2017.

25. Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Glass segmen-tation using intensity and spectral

polarization cues. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12622–12631, 2022.

26. 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.

27. Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16, pages 235–252. Springer, 2020.

28. Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pages 2160–2170, 2022.

29. 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.

30. 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.

31. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 2015.

32. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Con-volutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pages 234–241. Springer, 2015.

33. Karen Simonyan and Andrew Zisserman. Very deep convolu-tional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

34. 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.

35. Andrew Witkin. Scale-space filtering: A new approach to multi-scale description. In ICASSP'84. IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 9, pages 150–153. IEEE, 1984.

36. Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3907–3916, 2019.

37. Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16, pages 696–711. Springer, 2020.

38. Jinnan Yan, Trung-Nghia Le, Khanh-Duy Nguyen, Minh-Triet Tran, Thanh-Toan Do, and Tam V Nguyen. Mirrornet: Bio-inspired cam-ouflaged object segmentation. IEEE Access, 9:43290–43300, 2021.

39. Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1155–1162, 2013.

40. Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3166–3173, 2013.

41. Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camou-flaged object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4146–4155, 2021.

42. 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.

43. Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In Proceedings of the IEEE/CVF international conference on computer vision, pages 7234–7243, 2019.

44. Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 5217–5226, 2019.

45. Miao Zhang, Tingwei Liu, Yongri Piao, Shunyu Yao, and Huchuan Lu. Auto-msfnet: Search multi-scale fusion network for salient object detection. In Proceedings of the 29th ACM international conference on multimedia, pages 667–676, 2021.

46. 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.

47. 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.

48. Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3085–3094, 2019.

49. Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic polyp seg-mentation via multi-scale subtraction network. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, pages 120–130. Springer, 2021.

50. Yunfei Zheng, Xiongwei Zhang, Feng Wang, Tieyong Cao, Meng Sun, and Xiaobing Wang. Detection of people with camouflage pattern via dense deconvolution network. IEEE Signal Processing Letters, 26(1):29–33, 2018.

51. Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9141–9150, 2020.

52. Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In Proceedings of the European Conference on Computer Vision (ECCV), pages 121–136, 2018.