TOSN-Trans：Transparent Object Segmentation Network with Transformer

Tao Tao1; Jianfeng Yang1*; Jinsheng Xiao1; Wenfei Wu1
School of Electronic information, Wuhan University, Wuhan, China
Corresponding author: Jianfeng Yang. Email: yjf@whu.edu.cn

abstract

Due to the optical properties of glass materials, most glass appears transparent in RGB images. However, in depth images, different acquisition methods make glass visible. Therefore, Therefore, using RGB-D dual-channel feature input makes it easier to recognize and segment glass objects. Building on this concept, we propose a multi-layer symmetrical dual-channel network architecture, which can effectively realize trans-modal feature fusion of RGB-D images based on attention mechanism, and integrate Convolutional and Transformer architectures to extract local features and non-local dependencies, respectively. To further enhance segmentation accuracy and efficiency, this paper also designs a boundary optimization module. This module constructs a distance map based on edge prediction guidance, enabling high-precision glass edge recognition. To support this work, we collect a new dataset comprising 5551 sets of calibrated RGB-D images. The effectiveness and accuracy of the proposed glass segmentation method are rigorously evaluated quantitatively and qualitatively. The code for this paper has been published at: https://github.com/Jaccury/RGB-D-Transparent-object-segmentation.

Keywords: RGB-D, glass segmentation, feature fusion, dual-network, efficient transformer, boundary optimization

1. introduce

Transparent objects like glass allow most of the light incident on their surface to pass through and reflect only a minimal amount of light, resulting in a lack of prominent texture features in their image representation. In fact, these objects may exhibit very limited texture or even appear texture less, and in extreme cases, the image texture of glass may not be detectable at all (Yu R & Ren W,2024). Such transparent materials, represented by glass, defy the Lambertian assumption in optical behavior, which is the basis of 3D optical sensor operation, posing significant challenges to visual perception tasks. However, these perceptual scenarios are quite common, such as in automated production lines where vision-guided robotic arms locate and grasp transparent products like glass, household cleaning robots clearing transparent plastic bags, avoiding collision with broken glass containers, and in autonomous driving tasks that require avoiding glass walls commonly found in buildings. There is also a need for advanced visual assistance systems tailored for visually impaired individuals (Zhang B & Wang Z, 2023).

Due to the intricate optical characteristics of surface materials, the detection of glass using conventional visual sensors presents a notably challenging task compared to detecting opaque objects. Consequently, researchers have embarked on an exploration of multi-sensor information fusion, introducing depth sensors or thermal sensors as complementary tools for glass detection. However, thermal sensors are susceptible to environmental interference, while employing depth sensors to capture glass depth information is plagued by substantial detection errors due to the refractive and reflective nature of glass. These errors include missing depth measurements and inaccuracies in depth estimation (Cao J & Leng H, 2021). Collectively, these challenges contribute to the inherent unreliability of depth information in glass detection, thereby compromising the effectiveness of glass detection outcomes.

This study aims to analyze the challenges and difficulties faced by transparent object detection algorithms in real-world applications. We propose a method based on deep learning to construct an object detection network specifically tailored to address the challenges of false positives and false

negatives commonly encountered when detecting transparent objects against complex backgrounds. Finally, we evaluate the performance of the proposed network on a dataset of transparent object detection collected in complex environments.

To summarize, our main contributions are as follows:

(1) We conduct an in-depth analysis of the optical characteristics of glass and its pixel features in RGB images. Based on this analysis, we propose an image segmentation approach that leverages transparent body edge information supplemented by depth information, resulting in improved segmentation accuracy.

(2) We introduce a novel multi-layer symmetric dual-branch decoder-encoder architecture that effectively integrates image features and depth features, considering their differences and complementarity. In addition, we develop a new Efficient Transformer module to facilitate the cross-modal fusion of RGB-D image features. Furthermore, we present a boundary optimization module guided by edge information, enabling high-precision segmentation of glass edges.

(3) Our research effort also includes the collection of a new dataset comprising 5551 sets of calibrated RGB-D images captured by the RGB-D camera in real-world scenarios. The dataset includes manually annotated real segmentation masks, covering diverse daily life scenes.

2. related work

2.1. Transparent Object Recognition

In everyday scenarios, the visual appearance of glass undergoes considerable changes based on background settings and lighting conditions. Traditional algorithms for glass recognition often rely on intrinsic features or variations induced by glass itself, such as specular reflections, color similarities, or distortions in edge textures. Moreover, previous studies have integrated image features like HOG and SURF to investigate recognition and 3D reconstruction processes related to glass objects. The rapid evolution of neural networks has shifted conventional object recognition tasks towards deep learning paradigms. Consequently, deep neural networks have been increasingly applied to address the challenges associated with transparent object recognition, including glass. Lai et al. (LAI P-J & FUH C, 2015) mployed R-CNN methodologies to detect transparent objects within color images. They optimized the efficiency of selective search algorithms by incorporating features related to specular highlights and color similarities of transparent objects, thereby refining region proposals to exclude non-transparent areas. On a similar note, Khaing et al. (KHAING M P & MASAYUKI M, 2018) utilized SSD deep learning models to predict bounding boxes specifically tailored for transparent objects. Certain research (MEI H & YANG X, 2020) endeavors have delved into the reflective properties of glass mirrors. By contrasting features between the outer and inner aspects of mirrors, these studies aimed to detect glass objects by analyzing discontinuities in low-level color or texture features, particularly focusing on identifying mirror boundaries. However, it's worth noting that contrast features may not always be advantageous for transparent object segmentation, as semantic information within glass and low-level texture discontinuities may not be distinctly visible in transparent objects.

Conventional visual assistance systems integrate multi-sensor fusion methodologies, such as the fusion of RGB-D cameras with ultrasonic sensors, to overcome the complexities of handling transparent obstacles like glass objects, French windows, and glass doors. Chen et al. (CHEN H & WANG K, 2018) devised a multimodal stereo matching algorithm leveraging dual-depth sensors to refine depth measurement accuracy for transparent objects. Similarly, Yang et al. (YANG S-W & WANG C-C, 2018) developed a transparent object tracking framework utilizing data gathered from ultrasonic sensors and laser scanners, aiding mobile robots in effectively detecting and tracking transparent obstacles within indoor environments. However, it's worth noting that these approaches may lack the inclusion of visual sensor information, which is crucial for meeting the demands of transparent object recognition within visual assistance systems.

Furthermore, RGB-Depth (RGB-D) fusion techniques have been integrated into both traditional optimization methodologies and recent learning-based approaches. With the advent and refinement of Transformer architectures, algorithms for RGB-D feature fusion can be broadly classified into

three categories: purely convolution-based approaches (e.g., ESANet (Zhou J & Qian S, 2021), entirely Transformer architectures (e.g., VST (Liu ,2021) and Segmenter (R. Strudel, 2021), and hybrid networks combining convolutional networks with Transformers (e.g., CMX (Zhang J, 2023), EBS (J. Zhang,2021), DPANet (Chen Z,202), Segformer (E. Xie, 2021), among others).

2.2. Salient Object Detection

Salient Object Detection (SOD) aims to highlight prominent regions within an image, thereby reducing scene complexity and accurately capturing the essence of the image. This technique has found widespread application across various computer vision tasks and related domains. Traditional RGB-D-based SOD methods typically rely on graph-based approaches. Although these methods integrate depth features to enhance salient object detection tasks, they are limited by the design of manually crafted features that are based on a finite set of prior knowledge. This limitation constrains the method's ability to effectively represent low-level features, leading to significant biases in reasoning high-level features within complex scenes.

In recent years, there has been a surge of interest in deep learning-based SOD methods. Li et al. introduced LFSD (Li N, 2016), which employs machine learning techniques to validate the efficacy of incorporating light field data into SOD tasks. Additionally, Wang et al. (Wang T, 2020) proposed DLFS, followed by the emergence of several related methodologies such as Mo-SFM (Zhang M, 2019), LFNET (Zhang M,2020), Meanest (Jiang Y,2022), among others.

3. proposed method

3.1 RGB-D glass segmentation Network

Our segmentation network architecture is based on A standard decoder-encoder framework with skip-connections network architecture for dense segmentation tasks. In terms of the network as a whole, it is composed of two symmetric encoder branches, a decoder branch, and a trans-modal feature fusion module, as shown in Fig. 1.

As shown in Figure 1, the network proposed in this paper is a symmetric dual-branch encoder-decoder architecture. To be consistent, this paper denotes the output feature of Image branch at encoder component is expressed as $\{rai, i=1,2,3... n\}$, and denote the output features in the depth branch at the encoder component as $\{rbi, i=1,2,3... n\}$. Then, the encoders of the two branches gradually integrate multi-scale features. Finally, we aggregate the outputs of the dual-branch encoders and use the feature fusion module to generate the saliency map.

Specifically, in the encoder part, this paper is mainly based on two ResNet-50 backbone networks to accomplish the transformation of image features and depth features into feature tensor. This feature tensor is data in $H*W*C$ format, where H denotes height, W denotes width, and C denotes channel size. In order to fully and effectively extract the high-dimensional feature information of the two modalities, a symmetric four-layer dual-branch encoder structure is specially designed in this paper from the differences and complementarities of RGB images and depth images. In the part of trans-modal feature fusion, we input the result of this feature tensor supplemented with positional coding into the TFM module to generate feature fusion information in $H*W*C$ format. The fused features are input into the decoder, which has a four-layer structure, and gradually up-sampled to obtain the resolution of the input image. In particular, in order to further improve the reliability of this fusion feature, a boundary optimization module is designed in this paper to retrieve the glassy edges in the potential from the glassy optical properties to further improve the segmentation accuracy. Finally, the final segmentation mask is obtained based on the sigmoid function convolution layer.

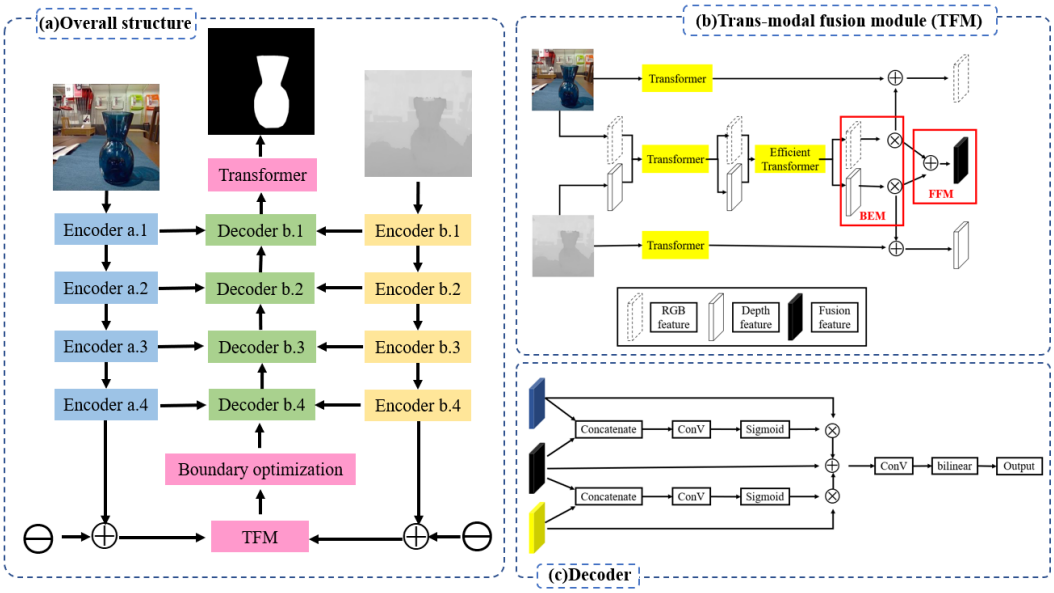


Fig.1. Architecture of transparent object segmentation network based on trans-modal feature fusion for RGB-D graphics. In this paper's network, the encoder is composed of two independent and symmetric ResNet-50 networks, which are mainly used to extract the high-dimensional features of the RGB image and depth image; and then after this paper proposes a kind of efficient transformer module to realize the feature fusion; and finally combines with the boundary optimization module to pass into the decoder, which produces the final segmentation result. The final segmentation result is generated.

3.2.1 Trans-modal fusion module (TFM)

The trans-modal feature fusion module in this paper is implemented based on Transformer, fully drawing on Transformer's attention mechanism. As mentioned earlier, image features F_i^R and depth features F_i^D both are $H*W*C$ structure type data. First, these two features are converted into 2*H*W*C structure feature diagram based on two corresponding transformer layers F_i^{RD}:

$$F_i^{RD} = transformer(stack(F_i^R, F_i^D)) \tag{1}$$

Then pass the F_i^{RD} passed into the Transformer Layer immediately following a sigmoid function to obtain a weight vector wi (data structure is $2*H*W*1$):

$$w_i = sigmoid(Linear(transformer(F_i^{RD}))) \tag{2}$$

In the feature fusion part, in particular, this paper proposes a lightweight feature fusion module to fuse features at different levels. The feature fusion module (FFM) in this paper is composed of two branches, both of which first process two cross-modal feature extraction modules (BEMs) for image features and depth features, respectively. Each branch starts with a 1×1 convolution for dimensionality reduction, which is used to reduce the number of parameters, followed by two residual blocks for nonlinear transformation of the two features for facilitating the subsequent feature fusion. Next, the model uses a 3×3 convolution to train and learn parameters that control the importance of each modal feature. Finally, context information is added to the fused features through a residual pooling. The 1×1 convolution and the 3×3 convolution other than the residual block play an important role in controlling the summation of the feature dimensions, as well as in adaptively fusing the features. Since depth features mainly play a supplementary role in semantic segmentation and the semantic segmentation network is more capable of recognizing Image features, the fusion

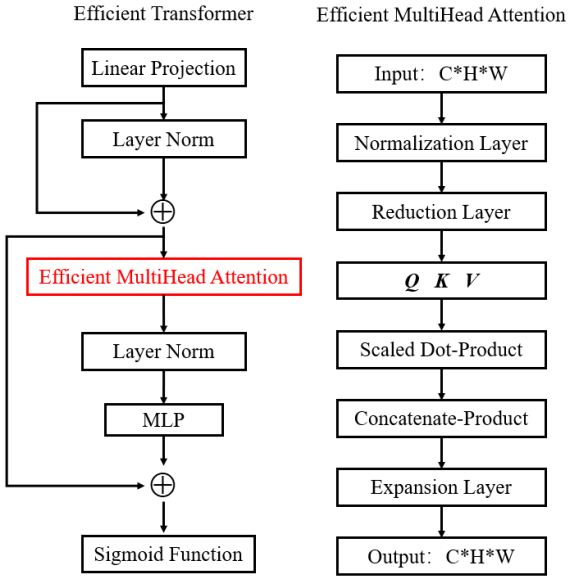
module is mainly used to reduce the effects of color blurring, illumination, etc. by complementing the Image features with depth features.

Of course, before feature fusion operation, it is also necessary to stack image features and depth features respectively:

$$F_{i+1}^{R} = transformer(F_i^{R}) + (w_i \otimes F_i^{RD})[:*H*W,:] \tag{3}$$

$$F_{i+1}^{D} = transformer(F_i^{D}) + (w_i \otimes F_i^{RD})[H*W*:,:] \tag{4}$$

As shown in Figure 1. (b), the TFM of this paper mainly contains four iterative blocks. This paper has made special treatment on the last Transformer Layer, mainly to solve the problems of insufficient ability of traditional CNN network to express global features and too much computation of Transformer.



(a) (b)

Fig.2. Efficient Transformer module structure diagram

In the TFM module, this paper only uses an Efficient Transformer put into the middle of the network as a capsule network, and optimizes the multi head attention in the classical Transformer. After the normalization layer, the Efficient Transformer inserts a reduction layer, which achieves the halving of the number of original feature channels, thus greatly reducing the computation of the network. Next, the linear layer projects the feature map into three matrices, *Queries*, *Keys*, and *Values*, according to the rules of the attention mechanism, and divides them into n segments, and then performs these three-matrix scaling dot product operations accordingly. Finally, learning the thinking of group convolution, it computes the merging of the large image blocks after splitting them into smaller ones and inserts the extension layer to restore to the original number of image channels. The output of the final ET O_i is as follows:

$$O_i = soft\max(\frac{Q_i K_i^{T}}{\sqrt{C}})V_i \tag{5}$$

Overall, in TFM, the two Transformer modules corresponding to the RGB image and the depth image extract non-local intra-modality dependencies from these two different modality features(F_i^R an F_i^D). Then extract the non-local modal dependencies between F_i^R andF_i^D in the stack based on the additional Transformer module of the intermediate branch. Considering the differences and complementarities of the two modes, a feature fusion network with excellent performance should ensure that the features between the two modes can complement each other. Therefore, instead of directly going along a certain dimension to split the cross-modal features

before inputting them into each modality as in traditional methods, this paper generates a weight vector w_i as a spatial attention mask to guide the discovery of positive features in the RGB image. Similarly, we cannot just hope that the spatial attention mask can completely uncover all the positive features in the RGB image, and we cannot guarantee the validity of each weight vector, so we need to combine the edge prediction of the RGB image and the edge guidance of the Depth image in a more organic way. This more complementary feature fusion is difficult to implement in the TFM module, so another boundary optimization module is designed in this paper.

3.2.2 boundary optimization

When image segmentation of transparent objects is performed, the pixel difference between the object and the background is small, and there is often some sense of boundary at the edge of transparent objects. In order to improve the quality of transparent object segmentation, this paper adds a branch in the image segmentation module to optimize the quality of our final segmentation.

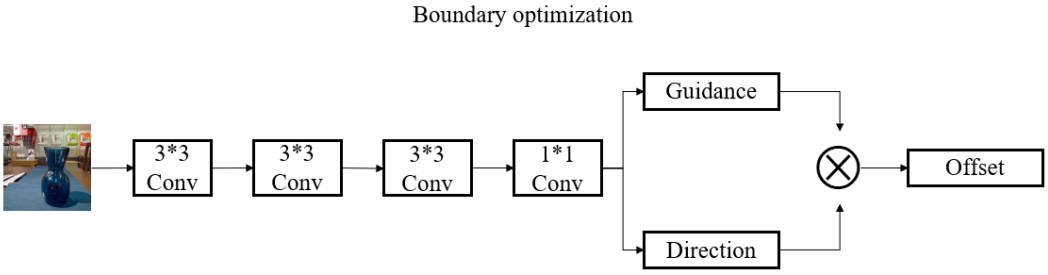


Fig.3. Network structure diagram of boundary optimization

The boundary optimization mainly consists of two steps. The first step is to predict and locate the boundary pixels, and the second step is to mark the corresponding depth information for each boundary pixel to confirm the internal pixels. The detailed architecture of the boundary optimization is shown in Figure 3. In order to fully consider the image features at the edge of the glass body, depth features are introduced as the guidance for edge detection, and such a dual branch network module is designed: Edge prediction branch based on Image features and edge guidance branch based on depth features. After the boundary guidance prediction is completed, the optimization results are applied to the semantic segmentation results using the offset coordinate offset branch.

Before the boundary guides prediction, a distance map needs to be generated. This distance map records the minimum Euclidean distance between pixels belonging to other categories, that is, the distance from the pixel to the edge. This is a mandatory step for edge detection for the boundary optimization module in this paper, which uses the binary cross entropy loss most edge distance map loss function to learn the edge detection rules.

$$Loss_1 = -\frac{1}{N}\sum_{i=1}^{N} y_i \bullet \log(P(y_i)) + (1 - y_i) \bullet \log(1 - P(y_i)) \tag{6}$$

In the boundary guided prediction process, the Sobel filter-based orientation points each pixel position to the pixel in its neighborhood that is farthest from the target boundary in the range [0°, 360°). Each pixel is then assigned to the corresponding category. During training, the discrete directions generated by this branch are supervised using a real distance map, and a multi-class cross-entropy loss is used as the direction prediction loss.

$$Loss_2 = -\frac{1}{m}\sum_{i=1}^{m} y_i \bullet \ln a + (1 - y_i) \bullet \ln(1 - a)) \tag{7}$$

The result of the final edge optimization, according to the optimization formula, the final boundary detection result with direction is obtained as follows.

$$\vec{S}_{pi} = S_{pi} + \Delta_{pi} \tag{8}$$

Among them. $\vec{S}_{pi}$ Is the redirection map, *pi* is the pixel coordinate, Δ_{pi} is the offset to the internal pixel.

3.2.3 Decoder

Corresponding to the encoder structure, the decoder in this paper also has a four-layer structure. Each layer receives features from image encoder and depth encoder respectively $(F_i^R \text{and} F_i^D))$, and the fusion features of the decoder at the layer before the reF_i^{RD}. Taking RGB image featuresF_i^Ras an example, tin this paper, we first combine $F_i^R \text{and} F_i^{RD}$, connecting three layers of convolution layers and a sigmoid function to calculate the weight volume of a single channel *W*. Finally, the output model of the decoder is:

$$Output_w = w_R \otimes F^R + w_D \otimes F^D + F^{RD} \tag{9}$$

In the same way as to the weight vector *wi* of TFM in Section 3.2.1, both w_R and w_D are spatial attention masks for feature fusion. However, in contrast, due to the memory pressure of high-resolution self-attention mechanism in solving intensive prediction tasks, this paper chooses the spatial attention scheme based on convolutional network in the decoder part. Meanwhile, in order to recover the spatial dimension, a convolutional layer and bilinear upsampling are also designed in each layer of decoder, as shown in Figure 1. (c).

4.Experiment

This algorithm has been implemented based on Python 3.8+Python 2.1.2+Ubantu 20.04 environment. The training network is ResNet-50 backbone network. The batch size is 8. Our model is trained with the AdamW optimizer for 300 epochs. The initialized learning rate is 10-4, and after 200 epochs it changes to 10-5.

4.1 RGB-D transparent object segmentation dataset

We captured the RGB-D images using a Femto Bolt TOF camera, which consists of a nearly juxtaposed RGB sensor and depth sensor. We used LabelMe to manually annotate the segmentation mask on the RGB images. Our new dataset covers a variety of scenes, such as libraries, shopping centers, galleries, train stations, museums, streets, factory floors, and houses, and generates 5551 RGB-D images from more than 40 scenes. Among them, we captured 370 pairs of scenes without glass in 7 scenes. To generate training and test segmentations, we randomly selected 23 scenes with glass and 5 scenes without glass for training and the others for testing.

4.2 Comparative experiments

4.2.1 Assessment of indicators

We use standard segmentation indicators: mean absolute error (MAE), intersection union set (I_{OU}), weighted F-measure (F_β), and balanced error rate (BER). The mean absolute error (MAE) is mainly used in foreground background segmentation tasks. It mainly calculates the average pixel error between predicted mask P and truth mask G. The definition is as follows:

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \left| P(i, j) - G(i, j) \right| \tag{10}$$

Where $P(i,j)$ denotes the prediction probability at the pixel position (i,j). F_β is the reconciled mean of average precision and average recall, defined as follows:

$$F_\beta = \frac{(1+\beta^2)(Precision \times Recall)}{\beta^2 Precision + Recall} \tag{11}$$

among β^2 =0.3. In this paper, we use the balanced error rate (BER) as an evaluation metric to consider the unbalanced region in the glass segmentation task, and provide a theoretical measure for evaluating the performance of the glass segmentation algorithm, defined as follows:

$$BER = \left(1 - \frac{1}{2}\left(\frac{TP}{Np} + \frac{TN}{Nn}\right)\right) \times 100 \quad (12)$$

Where, *TP*, *TN*, *Np*, and N_n represent the numbers of true positives, true negatives, glass pixels, and non-glass pixels respectively. These four metrics are commonly used in previous papers on glass segmentation, while only MAE is valid for evaluating images without glass. To evaluate the performance of those images without glass in the new dataset, we use the inverse intersection of concatenation (I_{OU}* and FPR) along with MAE. I_{OU}* takes the inverse mask and defines it. FPR is calculated as the ratio of the number of false positives (i.e., incorrectly detected glass) to the total number of images without glass.

4.2.2 quantitative analysis

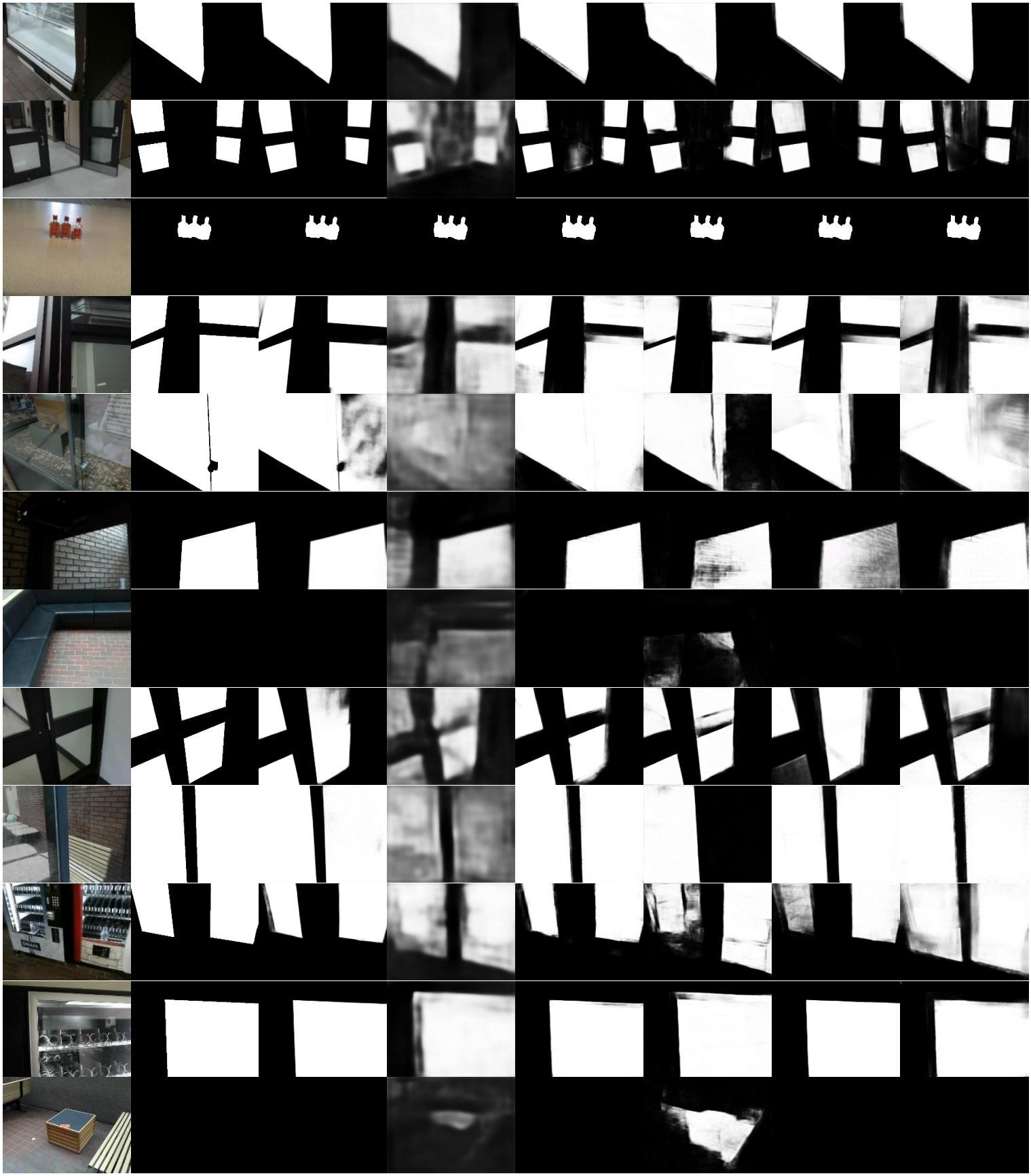
RGB-D image saliency detection is a new clue for glass segmentation. To demonstrate the versatility of our method on other RGB-D tasks, we retrain our model on 2500 training images in the dataset for RGB-D saliency target detection, and then evaluate it on 2500 test images in the dataset as well as 1000 images from the dataset. Comparing multiple state-of-the-art RGB-D SOD methods, we use MAE, IOU, Fβ, BER, and IOU* and FPR to evaluate the SOD results. We compare the algorithms in this paper with 30 state-of-the-art algorithms from other related fields in recent years, and the results of the comparison experiments are shown in Table 1. All these comparing methods use RGB-D images as inputs, trained and tested using the dataset of this paper.

4.2.3 Qualitative analysis

The results of the qualitative analysis are given in Figure 4. The method in this paper can accurately accomplish out the task of glass segmentation in various scenes in the dataset of this paper. Compared with the algorithm in this paper, the previous methods make a large number of errors in three aspects: (1) over-segmentation splitting the glass door and window openings into glass in some scenes; (2) fuzzy segmentation of the glass boundary; and (3) incomplete segmentation of the glass as a whole. In the figure, the scene without glass is also given, and the algorithm in this paper still performs well without misrecognizing the over-segmentation.

Tab. 1 Quantitative analysis result table of all methods

Methods	Glass-Images				Non-Glass-Images			All-Images
	MAE↓	Iou↑	F_β↑	BER↓	MAE↓	IOU*↑	FPR↓	MAE↓
DPANet	0.161	90.99	0.917	14.597	**0.003**	92.11	0.28	0.082
PSPNet	0.284	74.06	0.806	18.790	0.377	82.64	0.91	0.166
CMX	0.057	92.40	0.956	6.421	0.047	**99.70**	0.26	0.034
GDNet	0.163	77.63	0.837	15.620	0.418	79.21	0.77	0.122
TransLab	0.188	74.05	0.737	15.131	0.220	81.01	0.85	0.150
GlassNet	0.193	73.64	0.855	16.128	0.314	82.00	0.79	0.179
GlassSem	0.244	75.60	0.820	15.604	0.298	75.81	0.88	0.182
Segformer	0.061	89.82	0.934	7.702	0.163	92.79	0.43	0.135
Segmenter	0.217	78.53	0.812	15.418	0.303	82.60	0.61	0.127
SPNet	0.244	84.66	0.906	13.710	0.315	87.63	0.49	0.144
EBLNet	0.313	85.64	0.808	13.520	0.229	86.78	0.77	0.188
EBS [34]	0.309	88.44	0.851	9.885	0.301	94.07	0.27	0.220
ESANet [29]	0.315	89.97	0.939	6.865	0.311	87.55	0.321	0.042
VST [30]	0.148	89.16	0.944	8.977	0.326	84.11	0.45	0.149
DSCNet [65]	0.189	66.33	0.865	18.912	0.221	87.14	0.33	0.177
Ours	**0.032**	**92.78**	**0.961**	**4.491**	0.011	98.81	**0.12**	**0.031**



RGB ground truth ours DPANet CMX EBLNet EBS Segformer

Fig.4 Qualitative comparison experiment structure diagram. Qualitative comparison of our method with five most advanced methods

4.3 Ablation experiments

In order to verify the necessity of glass segmentation research based on RGB-D images, this paper designs two variant networks based on the previous network architecture, one based on RGB single channel features and the other based on depth single channel features. Similarly, similar to the experimental process in Section 6.1, the variant network training and testing are performed, and the results are shown in Table 2. This paper compares the method in this paper with four mainstream feature fusion modes, replacing MFM with simple feature summation (SFS), simple feature concatenation (SFC), per pixel attention fusion (PAF) and affine transformation (AT). The multi-mode fusion module used for RGB-D fusion in AFS-Net has also been used to replace our TFM. As shown in Table 3, the TFM module used in this paper has the best effect.

Tab.2 Feature Input Ablation Experiment Results Table

Methods	Glass-Images				Non-Glass-Images			All-Images
	MAE↓	Iou↑	F_β↑	BER↓	MAE↓	IOU*↑	FPR↓	MAE↓
RGB	0.120	83.94	0.919	8.294	0.193	94.49	0.65	0.180
Depth	0.394	61.91	0.638	19.011	0.497	86.41	0.13	0.372
RGB-D	**0.032**	**92.78**	**0.961**	**4.491**	**0.011**	**98.81**	**0.12**	**0.031**

Tab.3 Ablation Experiment Results Table

Methods	Glass-Images				Non-Glass-Images			All-Images
	MAE↓	Iou↑	F_β↑	BER↓	MAE↓	IOU*↑	FPR↓	MAE↓
SFS	0.162	89.34	0.926	7.600	0.304	91.43	0.51	0.104
SFC	0.141	90.14	0.945	7.190	0.258	93.10	0.44	0.077
PAF	0.122	91.99	0.948	7.113	0.197	94.66	0.31	0.070
AT	0.119	92.44	0.954	6.900	0.151	94.68	0.30	0.052
AFS	0.041	**93.09**	0.957	4.903	0.046	98.06	0.19	0.034
TFM	**0.032**	92.78	**0.961**	**4.491**	**0.011**	**98.81**	**0.12**	**0.031**

5. Conclusion

Transparent object detection is a major difficulty in the current object detection field. Based on the analysis of the optical properties of transparent objects, this paper proposes a high-precision glass RGB-D image segmentation algorithm based on the edge guidance module. This paper designs a multi-layer symmetrical dual-channel network architecture as a cross-modal fusion module to achieve feature fusion of RGB and depth. This module draws on the Transformer architecture, fully understands the differences and complementarities of cross-modal features, and generates a weight vector wi as a spatial attention mask to guide the discovery of positive features in RGB images, thereby achieving more efficient feature fusion. Furthermore, starting from the unilateral diffraction phenomenon at the edge of glass objects, this paper designs a boundary optimization module to focus on the boundary texture at the edge of transparent objects. This module first predicts and locates boundary pixels, then marks the corresponding depth information for each boundary pixel to confirm the internal pixels, and introduces depth information guidance to further improve the segmentation efficiency of transparent objects. In addition, we also collect a dataset of RGB-D glass scenes containing common life and work scenes. Our extensive evaluation shows that the performance of using RGB-D pairs is significantly better than that of using a single RGB image, and proves the superiority of our cross-modal fusion method over existing fusion methods using the same RGB-D input

Acknowledgements

This work was fully supported by Pingyang County, Zhejiang Province. Thanks to the relevant projects of the Science and Technology Bureau (No. 250071494) for funding.

References

Yu R, Ren W, Zhao M, et al. Transparent objects segmentation based on polarization imaging and deep learning[J]. Optics Communications, 2024, 555: 130246.doi: https://doi.org/10.1016/j.optcom.2023.130246.

Zhang B, Wang Z, Ling Y, et al. Shuffle Trans: Patch-wise weight shuffle for transparent object segmentation[J]. Neural Networks, 2023, 167: 199-212.doi: https://doi.org/10.1016/j.neunet.

Cao J, Leng H, Lischinski D, et al. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation[C]//Proceedings of the IEEE/CVF international conference on computer vision (ICCV), Montreal, QC, Canada, 2021, pp. 7068-7077, doi: 10.1109/ICCV48922.2021.00700.

LAI P-J, FUH C-S. Transparent object detection using regions with convolutional neural network[C]// IPPR conference on computer vision, graphics, and image processing. 2015, 2.

KHAING M P, MASAYUKI M. Transparent object detection using convolutional neural network[C]// International Conference on Big Data Analysis and Deep Learning Applications. Springer, 2018: 86-93.

CHEN H, WANG K, YANG K. Improving real sense by fusing color stereo vision and infrared stereo vision for the visually impaired[C]// Proceedings of the 2018 International Conference on Information Science and System. 142-146.

MEI H, YANG X, WANG Y, et al. Don't hit me! glass detection in real-world scenes[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3687-3696.

YANG S-W, WANG C-C. Dealing with laser scanner failure: Mirrors and windows[C]// 2018IEEE International Conference on Robotics and Automation. IEEE: 3009-3015.

Zhou J, Qian S, Yan Z, et al. ESA-Net: A network with efficient spatial attention for smoky vehicle detection[C]//2021 IEEE International Instrumentation and Measurement Technology Conference (I2MTC). IEEE, 2021: 1-6.

Liu, N. Zhang, K. Wan, L. Shao and J. Han, "Visual saliency transformer", Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), pp. 4722-4732, Oct. 2021.

R. Strudel, R. Garcia, I. Laptev and C. Schmid, "Segmenter: Transformer for semantic segmentation", Proc. IEEE/CVF Int. Conf. Comput. Vis., pp. 7262-7272, Oct. 2021.

Zhang J, Liu H, Yang K, et al. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers[J]. IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 12, pp. 14679-14694.

J. Zhang, J. Xie, N. Barnes and P. Li, "Learning generative vision transformer with energy-based latent space for saliency prediction", Proc. Adv. Neural Inf. Process. Syst., vol. 34, pp. 15448-15463, 2021.

Chen Z, Cong R, Xu Q, et al. DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection[J]. IEEE Transactions on Image Processing, 2020, 30: 7012-7024.

E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers", Proc. Adv. Neural Inf. Process. Sys. (NIPS), vol. 34, pp. 12077-12090, Dec. 2021.

Li N, Ye J, Ji Y, et al. Saliency Detection on Light Field[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016: 1605-1616.

Wang T, Piao Y, Lu H, et al. Deep Learning for Light Field Saliency Detection[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2020.

Zhang M, Li J, Wei J, et al. Memory-oriented decoder for light field salient object detection[C]. Proc of the 33rd International Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press,2019:898-908.

Zhang M, Ji W, Piao Y, et al. LFNet: Light field fusion network for salient object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 6276-6287.

Jiang Y, Zhang W, Fu K, et al. MEANet: Multi-modal edge-aware network for light field salient object detection[J]. Neurocomputing, 2022, 491: 78-90.