TROSD: A New RGB-D Dataset for Transparent and Reflective Object Segmentation in Practice

Tianyu Sun, Guodong Zhang, Wenming Yang, *Senior Member, IEEE,* Jing-Hao Xue, *Senior Member, IEEE,* Guijin Wang, *Senior Member, IEEE*

***Abstract*—Transparent and reflective objects are omnipresent in our daily life, but their unique visual and optical characteristics are notoriously challenging even for state-of-the-art deep networks of semantic segmentation. To alleviate this challenge, we construct a new large-scale real-world RGB-D dataset called TROSD, which is more comprehensive than existing datasets for transparent and reflective object segmentation. Our TROSD dataset contains 11,060 RGB-D images with three semantic classes in terms of transparent objects, reflective objects, and others, covering a variety of daily scenes. Together with the dataset, we also introduce a novel network (TROSNet) as a high-standard baseline to assist other researchers to develop and benchmark their algorithms of transparent and reflective object segmentation. Moreover, extensive experiments also clearly show that the proposed TROSD dataset has an excellent capacity to facilitate the development of semantic segmentation algorithms with strong generalizability.**

***Index Terms*—RGB-D Dataset, Transparent and Reflective Object, Semantic Segmentation.**

I. Introduction

SEMANTIC segmentation is a fundamental task in computer vision. Transparent and reflective objects are omnipresent in our daily life. However, it remains notoriously challenging, even for state-of-the-art deep networks, to attain accurate semantic segmentation of transparent and reflective objects. This is mainly due to the unique visual and optical characteristics of such objects, which are extremely distinct from most types of objects in popular large-scale datasets for semantic segmentation. The aim of this paper is to alleviate this challenge.

There are two major obstacles to achieving our aim.

First, a proper dataset for training and testing semantic segmentation algorithms, as we all know, is of the same importance as the segmentation algorithms themselves [1], [2]. However, unfortunately, existing segmentation datasets seldom contain images of transparent or reflective objects, causing the trained models lacking generalizability on such objects. In the past few years, to tackle this obstacle, Yang *et al.* [3] provide a large-scale RGB mirror segmentation dataset (MSD); Mei *et al.* [4] construct an RGB glass detection dataset (GDD) to segment glass; and Seib *et al.* [5] provide a depth dataset with transparent drinking glasses. Some other researchers choose to make the concession by resorting to synthetic RGB-D datasets. Sajjan *et al.* [6] construct a large-scale synthetic RGB-D dataset to segment transparent objects. However, none of these datasets contains real-world RGB-D images of both transparent and reflective objects, making it hard for researchers to train and test segmentation models on real-world scenes where either or both of these two types of objects can be present.

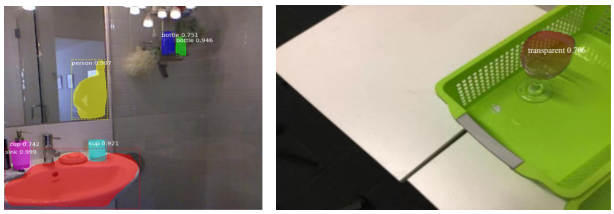

Fig. 1. Mis-segmentation by (left) Mask R-CNN [7] and (right) TransLab [8].

Second, most existing semantic segmentation algorithms do not appreciate transparent and reflective objects, hence the presence of such objects in a scene can significantly degrade the performance of these algorithms. For instance, when there are reflective objects in the scene, semantic segmentation algorithms can falsely recognize the virtual image as a real object. As shown in Fig. 1, the state-of-the-art Mask R-CNN [7] is misled by the reflected visual texture (on the left). As for transparent objects, relative low reflectivity and ambiguous edge make it hard to be observed [9]. On the right of Fig. 1, TransLab [8] fails to segment the base of the goblet. Consequently, these will degrade the performance in various 3D computer vision tasks (*e.g.,* 3D reconstruction, depth estimation).

Therefore, our research objective is to address these two obstacles. To this end and to promote further research in this area, we construct a large-scale real-world RGB-D dataset, as well as a high-standard baseline method, for semantic segmentation of transparent and reflective objects.

The novelties and contributions of this paper are summarized as follows.

First, we construct a new large-scale transparent and reflective object segmentation dataset (TROSD), which contains 11,060 real-world RGB-D images with detailed annotations

T. Sun is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.

G. Zhang and W. Yang are with the Shenzhen International Graduate School/Department of Electronic Engineering, Tsinghua University, Shenzhen 518055, China.

J.-H. Xue is with the Department of Statistical Science, University College London, London WC1E 6BT, U.K.

G. Wang is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, and with Shanghai Artificial Intelligence Laboratory.

of transparent and reflective objects (*e.g.*, glasses, plastics, mirrors). We believe TROSD will substantially assist other researchers to make further improvements for semantic segmentation in future practice.

Second, we introduce a new semantic segmentation network (TROSNet) as a high-standard baseline for segmenting transparent and reflective objects. TROSNet pays particular attention to low-level features, to identify and preserve content discontinuity at the boundaries of transparent and reflective objects. It also exploits the best of both RGB and depth features through multi-modal fusion.

Last but not least, extensive experiments verify the validity and generalization capability of the proposed TROSD dataset and the superiority of TROSNet to many other state-of-the-art semantic segmentation methods.

II. Related work

A. Semantic segmentation

Semantic segmentation is a fundamental computer vision task and has many potential applications [10]. In the past few years, semantic segmentation has achieved significant success based on fully convolution networks (FCNs) [11]. Some methods (*e.g.*, PSPNet [12], RefineNet [13] and DeepLab [14]) aggregate different region based contexts. Others (*e.g.*, OCNet [15] and ANNNet [16]) introduce a non-local block to leverage the long-range dependencies via attention mechanism. These algorithms are mostly trained and tested on RGB datasets such as PASCAL VOC (RGB) [17], Cityscapes (RGB) [1] and SUN RGB-D (RGB-D) [2].

With the emergence of RGB-D sensors, researchers turn to investigating how to assist image processing with 3D information [18]–[21], including depth images and point cloud. Wang *et al.* [22] introduce D-CNN and average pooling to improve the capability of handling geometric information from depth images. Jiao *et al.* [23] use a distilling geometry-aware embedding method to exploit the helpful depth information. Zhang *et al.* [24] propose a novel PAP framework to predict depth, surface normal and semantic segmentation. Qi *et al.* [25] design PointNet, a unified architecture for applications including part segmentation and scene semantic parsing. Fan *et al.* [26] propose PST convolution to achieve informative representations of point cloud sequences and improve semantic segmentation. Thomas *et al.* [27] present KPConv, which operates on point clouds without intermediate representation. However, as shown in Fig. 1, directly applying these algorithms to transparent and reflective object segmentation can be problematic.

B. Transparent and reflective object segmentation datasets

In recent years, researchers pay much attention to transparent and reflective object segmentation, but one of the main obstacles is the lack of relative datasets. Therefore, some researchers work to construct segmentation datasets for certain types of transparent or reflective objects, such as mirrors and glasses.

Xie *et al.* [8] build Trans10k-v2, a large-scale dataset of 10,428 images of glass and corresponding masks for segmenting glass from a single RGB image. Yang *et al.* [3] introduce a large-scale dataset named MSD (RGB-D) of 4,018 RGB-D images with mirrors and corresponding masks. Seib *et al.* [5] provide an RGB-D dataset of 440 images with depth information for four different types of transparent drinking glasses. The ClearGrasp dataset (RGB-D) [6] is a large-scale synthetic RGB-D dataset with over 50,000 synthetic images and 286 real-world images. However, these works only focus on limited types of objects, either transparent or reflective, instead of building a comprehensive dataset for segmenting both transparent and reflective objects.

Different from past works, we construct a segmentation dataset that contains both transparent and reflective objects (*e.g.*, mirrors, drinking glasses, transparent plastic), which is more general than other datasets mentioned above.

C. Transparent and reflective object segmentation methods

Recently, more and more researchers are working on transparent and reflective object segmentation. Yang *et al.* [3] propose a binary classifier MirrorNet (RGB) for mirror segmentation from RGB images, which achieves a good performance on the MSD dataset. Li *et al.* [28] introduce Mirror-YOLO (RGB) with an outstanding result on MSD. Seib *et al.* [5] propose an RGB network to detect the existence of transparent drinking glasses via exploiting undefined values in the depth images. ClearGrasp (RGB) [6] uses deep convolution networks to segment transparent objects from RGB images for robotic manipulation with synthetic training data. Mei *et al.* [4] explore abundant contextual cues for robust glass segmentation with a novel large-field contextual feature integration method (RGB). Mei *et al.* [29] introduce a novel mirror segmentation method that leverages both RGB and depth information obtained by ToF-based cameras. However, most of these methods only use a single modality (RGB or depth) to detect reflective or transparent objects. What is more, none of these works builds a unified framework for segmenting both transparent objects and reflective objects.

In this context, to offer a high-standard baseline, we introduce a novel multi-modal method for segmenting both transparent objects and reflective objects. This baseline method exploits both depth distribution and RGB visual appearance.

III. TROSD: A new transparent and reflective object segmentation dataset

We construct a large-scale dataset termed TROSD, which contains 11,060 RGB-D images with transparent objects and reflective objects. In the following subsections, we shall detail the construction and analysis of TROSD [1].

A. Dataset construction

TROSD is an RGB-D image dataset for segmenting both transparent and reflective objects. As shown in Table I, TROSD consists of images from three sources: SUN RGB-D [2], ClearGrasp [6] and new data captured by ourselves.

[1]TROSD is available at http://www.tsinghua-ieit.com/trosd.

TABLE I
COMPOSITION OF THE PROPOSED RGB-D DATASET TROSD. TROSD CONTAINS SOME RGB-D IMAGES OF TRANSPARENT AND REFLECTIVE OBJECTS FROM TWO EXISTING RGB-D DATASETS AND THE IMAGES COLLECTED BY OURSELVES.

Part	Source	Number	
		Train	Test
1	SUN RGB-D [2]	673	455
2	ClearGrasp (real) [6]	-	752
3	Our new data	6,748	2,432
Total		7,421	3,639

The new data captured by ourselves compose the majority in TROSD, making up more than 80 percent of all data. We capture RGB-D images that contain different kinds of transparent objects (*e.g.*, glass bottles, plastic bottles, windows) and reflective objects (*e.g.*, mirrors, metals). In total there are more than 50 different types of transparent and reflective objects in the TROSD dataset.

For new data collection, we use an iPad 2018 equipped with a Structure Sensor [30] to capture RGB-D images (640×480) of transparent and reflective objects. Structure Sensor is an infrared structured light sensor. It has one infrared sender and one infrared receiver. The sensor projects an IR speckle pattern to the target object. This pattern is then reflected back to the sensor, and the depth is calculated based on the time between sending and receiving. The depth image is sent back to iPad and processed to match the RGB image pixel-by-pixel.

For object selection, we consider common items (mainly including glass products, mirrors, plastic and metal) in different sizes, shapes or colors. As for scenes, the proposed TROSD dataset contains 14 different scenes (*e.g.*, living room, bathroom, office) in total.

For the processing of new data, we first use a calibration procedure [31] to obtain camera parameters for both RGB and depth sensors and align the RGB images with depth images. Then, we split the dataset into a training set and a test set. Moreover, we make sure that the scenes in the training set do not appear in the test set. As a result, 6,748 RGB-D images belong to the training set and other 2,432 images belong to the test set. Finally, we annotate these images manually with LabelMe (an image annotation tool) [32].

Additionally, as shown in Table I, we include some RGB-D images from the existing RGB-D recognition benchmark datasets (SUN RGB-D [2] and ClearGrasp [6]) into TROSD. There are images consisting of transparent or reflective objects in the SUN RGB-D dataset, but it does not provide annotations for these objects. In this case, we manually annotate the masks of these objects in totally 1,128 RGB-D images and resize these images to the size of 640×480. Afterwards, we collect the real test data from the ClearGrasp dataset [6] with annotated masks and generate 752 RGB-D images with data augmentation. For all the depth images, we calibrate the depth values to the same size and scale, projecting them to the range of 0 to 255 with linear projection.

Some example images from our TROSD dataset are shown in Figs. 6-8 in Section III-C.

B. Dataset analysis

This section presents a comprehensive analysis of the TROSD dataset.

TABLE II
COMPARISON OF TROSD WITH EXISTING TRANSPARENT AND REFLECTIVE OBJECT DATASETS. TROSD CONTAINS MORE DIVERSE OBJECTS AND REAL RGB-D IMAGES.

Dataset	Modalities	Images	Objects
GDD [5]	RGB	3,900	transparent
GSD(2021) [33]	RGB	4,102	transparent
Trans10k-v2 [8]	RGB	10,428	transparent
ClearGrasp(real) [6]	**RGB-D**	286	transparent
GSD(2022) [34]	**RGB-D**	3,009	transparent
MSD [3]	RGB	4,018	reflective
PMD [35]	RGB	6,461	reflective
RGBD-Mirror [22]	**RGB-D**	3,049	reflective
Mirror3D [36]	**RGB-D**	5,894	reflective
TROSD	**RGB-D**	**11,060**	**both**

We first provide a summary comparison of TROSD and other benchmark datasets in Table II. The proposed TROSD dataset offers three advantages: 1) it considers both transparent objects and reflective objects at the same time; 2) it provides two modalities of both RGB and depth images, and the depth channel can bring more geometrical information that can promote the development of segmentation networks; and 3) it contains many more images, especially RGB-D images. In short, the proposed TROSD dataset is more comprehensive than existing transparent and reflective object datasets.

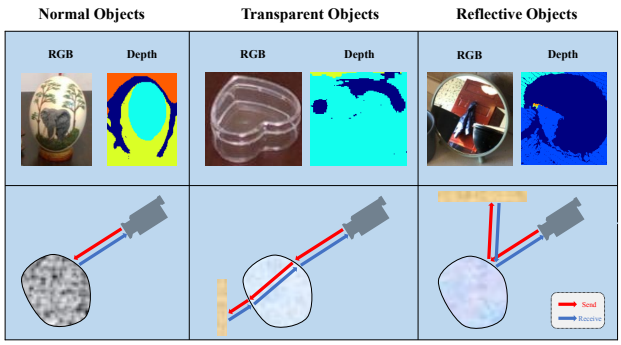

Fig. 2. Depth calculation of different objects.

We then illustrate some insights of our dataset. Due to special optical characteristics of transparent objects and reflective objects, researchers tend to seek more spatial information from depth images [37], [38]. However, the depth calculation of RGB-D camera is also, to some extent, affected by such optical characteristics. As shown in Fig. 2, reflective objects reflect IR lasers to surrounding objects and transparent objects are penetrated by IR lasers. It means that these objects also produce noise in the depth images and generate random artifacts. In this case, fusing RGB and depth channels can better address the problem. Our TROSD can provide RGB and depth images for transparent objects and reflective objects, enabling

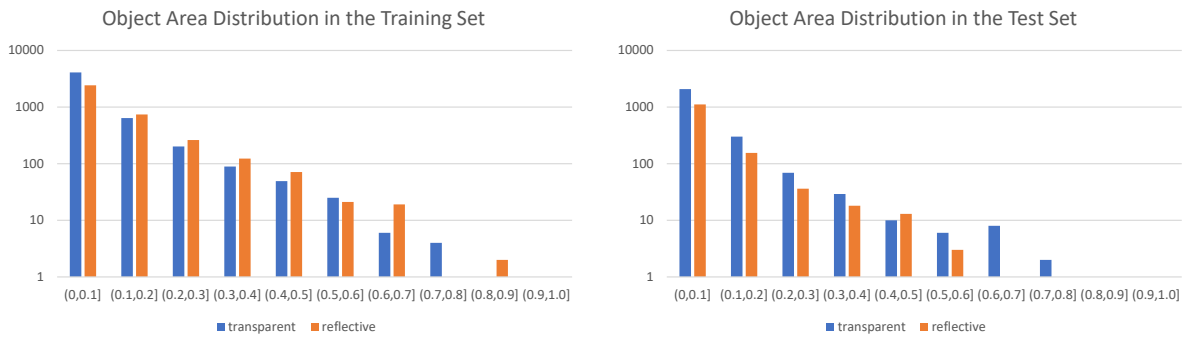

Fig. 3. Area distribution for transparent and reflective objects in the training and test sets.

researchers to acquire semantic and spatial information for the segmentation of such objects.

Then we illustrate three summary statistics of our new dataset:

Mask area distribution Fig. 3 shows the distribution of area proportion (from 0 to 1) of masks for transparent and reflective objects in each image. The highly skewed distribution indicates that, in most of the images, transparent objects and reflective objects only occupy a small area. It means that our dataset could comprehensively test the segmentation methods on small transparent and reflective objects.

TABLE III
AN OVERVIEW OF THE NUMBER OF CLASSES OF OBJECTS IN ALL IMAGES IN OUR DATASET.

Part	Objects included	Number	
		Train	Test
1	Only transparent	3,481	2,307
2	Only reflective	2,040	1,136
3	Both	1,606	196
4	None	294	0
Total		7,421	3,639

Semantic classes distribution Table III shows the number of images with different types of objects in our dataset: with only reflective objects, with only transparent objects, with both transparent and reflective objects, and with background only. We can see that our dataset contains over 1,800 images with both transparent objects and reflective objects simultaneously. In addition, we include images with neither transparent nor reflective objects to introduce more diversity to our dataset.

Object location distribution Fig. 4 is the rendered heatmap of object location, which suggests that the transparent and reflective objects are largely located in the middle area of the images.

C. Example images from TROSD

In this section we show some examples from our TROSD. Each example is presented with RGB image, depth image and mask ground truth. RGB images are shown directly in the size of 640×480. Depth images are rendered in the scale shown in Fig. 5, with colors in the left of the bar standing for smaller

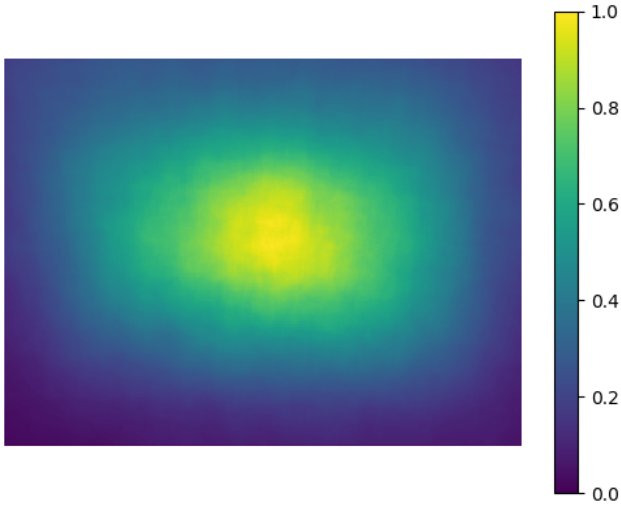

Fig. 4. Location heatmap of transparent and reflective objects.

depth and colors in the right for larger depth. The depth images are also visualized in the size of 640×480. The ground-truth masks are piled upon RGB images; in the ground-truth masks, the red pixels denote transparent objects, and the green pixels denote reflective objects.

Fig. 5. Scale of the depth image.

We show these examples in three figures, which include images with only transparent objects, with only reflective objects and with both transparent and reflective objects, respectively.

Fig. 6 shows four examples containing only transparent objects. The transparent objects we take into consideration include drinking glasses, glass tables, windows, *etc*. Some images also contain multiple transparent objects.

Fig. 7 shows four examples containing only reflective objects. The reflective objects we place in the scene include mirrors and mobile screens.

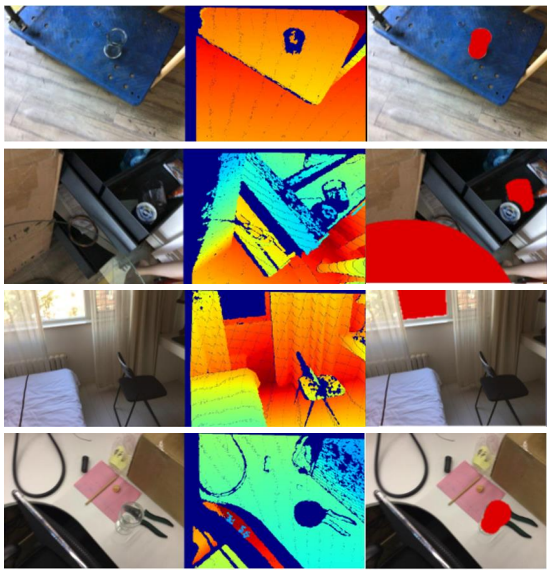

Fig. 6. Examples of images with only transparent objects.

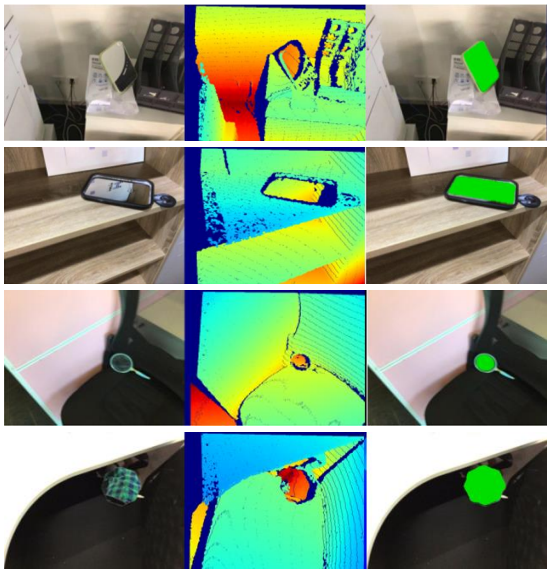

Fig. 7. Examples of images with only reflective objects.

Fig. 8 shows two examples containing both transparent and reflective objects. In these examples, we put together a drinking glass and a mirror. To make the scenes more complicated, we place these objects close enough to block each other's view.

IV. TROSNet: A high-standard baseline for transparent and reflective object segmentation

We now presents TROSNet, a new high-standard baseline for transparent and reflective object segmentation. TROSNet exploits multi-modal features of both the textures from RGB images and the 3D geometric discontinuities from depth images.

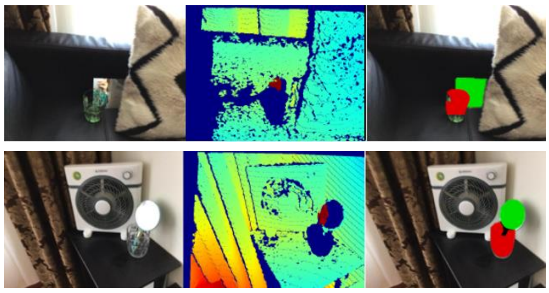

Fig. 8. Examples of images with both transparent and reflective objects.

A. Overview

TROSNet consists of four parts (Fig. 9): Encoder, Feature Fusion, Boundary Refinement and Decoder.

Encoder In the encoder, we feed the input RGB (I) and depth (D) images into a two-stream backbone (*e.g.*, the encoder part of a ResNet [39] consisting of *Conv1*, *Layer1*, *Layer2*, *Layer3* and *Layer4* in Fig. 9), to extract useful RGB features ($F_{rgb} \in \mathcal{R}^{C\times H\times W}$) and depth features ($F_d \in \mathcal{R}^{C\times H\times W}$) for semantic segmentation. Here C, H and W denote the channel number, spatial height and width, respectively. We also design a Cascaded Multi-modal Fusion (CMF) unit (Fig. 10) to preserve and enhance discontinuity details at the outlines of transparent and reflective objects.

Feature fusion This module aims to enhance RGB and depth information from the encoder by fusing the enhanced RGB feature of $\overline{F}_{rgb}$ with the enhanced depth feature of $\overline{F}_d$.

Boundary refinement Considering that some researches on segmentation focus on boundary learning to acquire the exact shape of the target [33], [40], [41], we apply residual learning [39] (BR block in Fig. 12) to refine the boundary information in the feature maps. Considering the various types of transparent and reflective objects, we implement the Adaptive Layer-Instance Normalization (AdaLIN) [42], an outstanding feature-extraction network robust to the shape and texture of the image.

Decoder The encoded features from the encoder are computed with output stride16. In the decoder (Fig. 9), we insert several 3×3 convolution layers to refine the features followed by another simple bilinear upsampling by a factor of two. After four times of upsampling, we attain the final predicted semantic map.

B. Cascaded multi-modal fusion unit in the encoder

As shown in Fig. 10, the input of the CMF unit contains raw RGB-D features and low-level complementary features ($\mathcal{A}$) (i.e., the complementary information between F_{rgb} and F_d at a lower level).

We design the novel CMF unit, aiming at extracting more useful and detailed features from the input. In addition, unlike

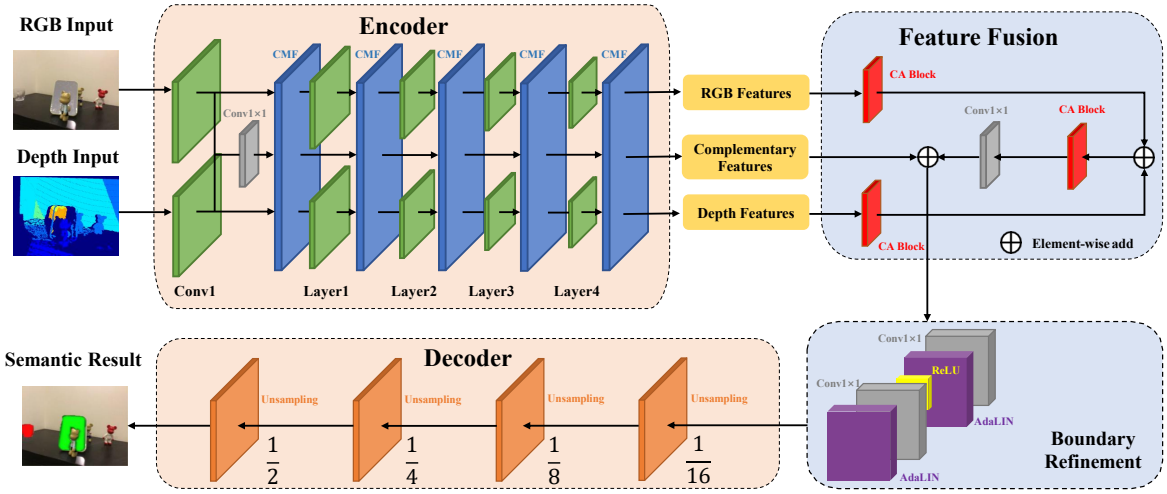

Fig. 9. The overall network structure of TROSNet. Firstly, given an input RGB-D image, we use a two-stream encoder to extract features from RGB and depth images, respectively. Then, we use a Feature Fusion module to fuse RGB and depth features. Next, a Boundary Refinement module is used to refine the boundary details of the features. Finally, a decoder predicts the semantic label maps.

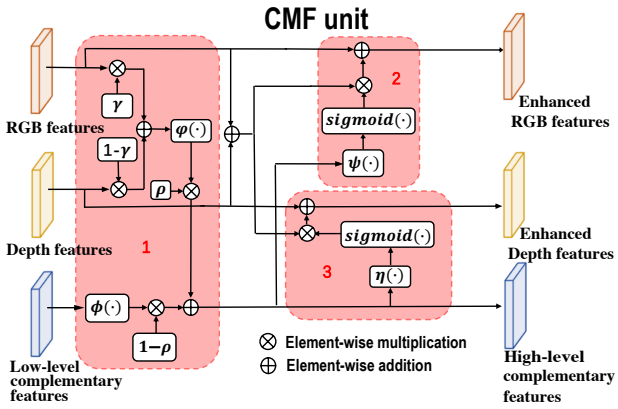

Fig. 10. Cascaded Multi-modal Fusion Unit.

treating the depth input simply as an additional fourth channel, we highlight its significance through our network.

In this case, we design the CMF unit based on the following two principles: 1) to preserve low-level features as much as possible; and 2) to enhance RGB-D and complementary features effectively.

To preserve low-level features, the proposed CMF unit implements a skip-connection path between the two adjacent blocks of the encoder. In this way, it provides useful low-level details to the high-level RGB-D features and is helpful to identify content discontinuity at the boundaries of these objects. In detail, to preserve low-level features between the two adjacent blocks of the encoder, the CMF unit enhances the low-level complementary features ($\mathcal{A}$) as

$$\overline{\mathcal{A}} = (1-\rho)\phi(\mathcal{A}) + \rho\varphi(\gamma F_{rgb} + (1-\gamma)F_d), \quad (1)$$

where ρ and γ are parameters learned by the network, and $\phi(\cdot)$ and $\varphi(\cdot)$ are two transformation functions which can adaptively transform raw feature maps to different embeddings. As given in Eq.(1), to calculate the enhanced complementary features ($\overline{\mathcal{A}}$), the CMF unit takes advantage of both current RGB-D features (F_{rgb} and F_d) and complementary features ($\mathcal{A}$). The learnable parameters ρ and γ can combine current RGB-D features and complementary features selectively to generate the enhanced complementary features ($\overline{\mathcal{A}}$). Then, the enhanced complementary features are fed into the next stage of the encoder via a skip-connection path to preserve low-level features.

To obtain the enhanced RGB-D discontinuity features ($\overline{F}_{rgb}$ and $\overline{F}_d$) (part two and part three in Fig. 10) with the enhanced complementary features ($\overline{\mathcal{A}}$) (part one in Fig. 10), the CMF unit enhances F_{rgb} and F_d and takes advantages of both the precedent complementary features and the current RGB-D features. This process of complementary information enhancement for F_{rgb} and F_d can be expressed as

$$\overline{F}_{rgb} = \mathrm{sigmoid}(\psi(\overline{\mathcal{A}}))(F_{rgb} + F_d) + F_{rgb}, \quad (2)$$

$$\overline{F}_d = \mathrm{sigmoid}(\eta(\overline{\mathcal{A}}))(F_{rgb} + F_d) + F_d, \quad (3)$$

where $\mathrm{sigmoid}(\cdot)$ is the activation function, and $\psi(\cdot)$ and $\eta(\cdot)$ are two transform functions. In this way, the RGB and depth features can accumulate useful information from complementary modality. In addition, this process is also guided by the precedent complementary features ($\mathcal{A}$), which means that the enhanced RGB and depth features also benefit from precedent low-level RGB and depth features.

C. Feature fusion

Considering that channel-wise attention can enhance modal information and channel information in multi-modal problems [43], [44], we design our module with an effective multi-modal fusion strategy based on channel-wise attention

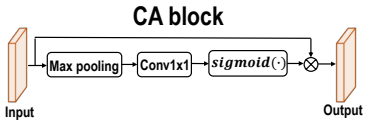

Fig. 11. Channel-wise Attention Block.

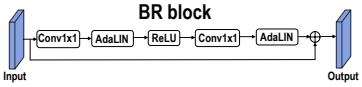

Fig. 12. Boundary Refinement Block.

mechanism to fully utilize semantic information of the RGB and depth images, as shown in Fig. 9. In this Feature Fusion stage, RGB and depth features are firstly enhanced via the Channel-wise Attention block (CA block in Fig. 11) [45] respectively. The CA block aggregates contrasted features via max-pooling and adopts convolution operations to generate the attention weights. Then, the features are adaptively enhanced based on the attention weights. Finally, we also merge the final complementary features with the fused features via element-wise addition to preserve low-level features (Fig. 9).

D. Implementation details

In our TROSNet, the two-stream encoder is based on the ResNet-50 [39] pre-trained on the ImageNet dataset [54]. As for the CMF unit in Fig. 10, $\phi(\cdot)$, $\varphi(\cdot)$, $\psi(\cdot)$ and $\eta(\cdot)$ are four 1×1 convolution layers, which can adaptively transform the raw features to different embeddings. Parameters ρ and γ are initialized to 0.5. After the Feature Fusion, we cascade four Boundary Refinement blocks (Fig. 12) to refine the boundary details of the feature maps. Eventually, we exploit a multi-level output based decoder to facilitate the learning process.

To achieve better performance, following Yang *et al.* [3], we utilize a multi-scale loss for optimizing our network. The overall loss function is

$$Loss = \sum_{t=0}^{T} \eta_t L_t, \tag{4}$$

where L_t is the Cross Entropy loss between the t-th level predicted segmentation map and the ground truth that is downsampled by a factor of 2^t; and η_t is the balancing weight of L_t. In our network, there are four loss terms in Eq.(4) ($T = 3$) and all η_t are empirically set to 1.

For training, we use Stochastic Gradient Descent (SGD) optimizer. Momentum and weight decay are set to 0.9 and 0.0001, respectively. The batch size is 16. As for learning rate, we refer to a mix of trigonometric function and exponential function which could accelerate our training. The training would reach the convergence of loss function between 100 to 200 epochs and, to play safe, we set the epoch number as 300.

For data augmentation, input and target images are horizontally flipped with probability 0.5, scaled with $s \in [1, 1.5]$, and rotated by $r \in [-5, 5]$ degrees on training. Finally, we crop the input image to the size of 640×480.

V. Experimental Results

A. Evaluation metrics

Three common metrics for semantic segmentation are adopted here for performance evaluation: mean pixel accuracy of different categories (mAcc), Intersection-over-Union of different categories (IoU), and mean IoU (mIoU).

B. Results on the TROSD dataset

We first compare our TROSNet with state-of-the-art methods on the TROSD dataset. All the competitors are listed in Table IV with their best results, by using their public source codes and under the same data augmentation strategy.

As shown in Table IV, our TROSNet outperforms the state-of-the-art methods, achieving the best performance in terms of mIoU (72.01%) and mAcc (81.21%). Especially, as for mean

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON TROSD. R: REFLECTIVE OBJECTS. T: TRANSPARENT OBJECTS. B: BACKGROUND. THE BEST RESULTS ARE IN **bold**. * FOR EBLNET, WE IMPLEMENT RESNET-50 AND RESNEXT-101 AS BACKBONE FOR TRANSPARENT OBJECTS AND REFLECTIVE OBJECTS SEPARATELY.

Method	Input	Backbone	IoU (%) R	IoU (%) T	IoU (%) B	mIoU (%)	mAcc (%)
RefineNet [13]	RGB	ResNet-101	21.32	37.32	92.37	50.34	63.59
ANNNet [16]	RGB	ResNet-101	22.31	41.30	93.43	52.35	62.49
Trans4Trans [46]	RGB	PVT [47]	27.69	39.22	94.16	53.69	61.82
PSPNet [12]	RGB	ResNet-101	26.35	44.38	94.19	54.97	64.14
OCNet [15]	RGB	ResNet-101	31.76	46.52	95.05	57.78	64.46
TransLab [8]	RGB	ResNet-50	42.57	50.72	96.01	63.11	68.72
DANet [48]	RGB	ResNet-101	42.76	54.39	95.88	64.34	70.95
SSMA [49]	RGB-D	ResNet-50	24.70	29.04	89.98	47.91	67.72
FRNet [50]	RGB-D	ResNet-34	28.37	36.59	92.18	52.38	63.94
EMSANet [51]	RGB-D	ResNet-101	27.53	44.10	96.14	55.92	71.63
FuseNet [52]	RGB-D	VGG-16	37.30	43.29	94.97	58.52	66.13
RedNet [53]	RGB-D	ResNet-50	48.27	47.57	95.76	63.87	69.23
EBLNet [40]	RGB-D	ResNet*	51.75	50.12	94.57	65.49	67.39
TROSNet	RGB-D	ResNet-50	**62.27**	**57.23**	**96.52**	**72.01**	**81.21**

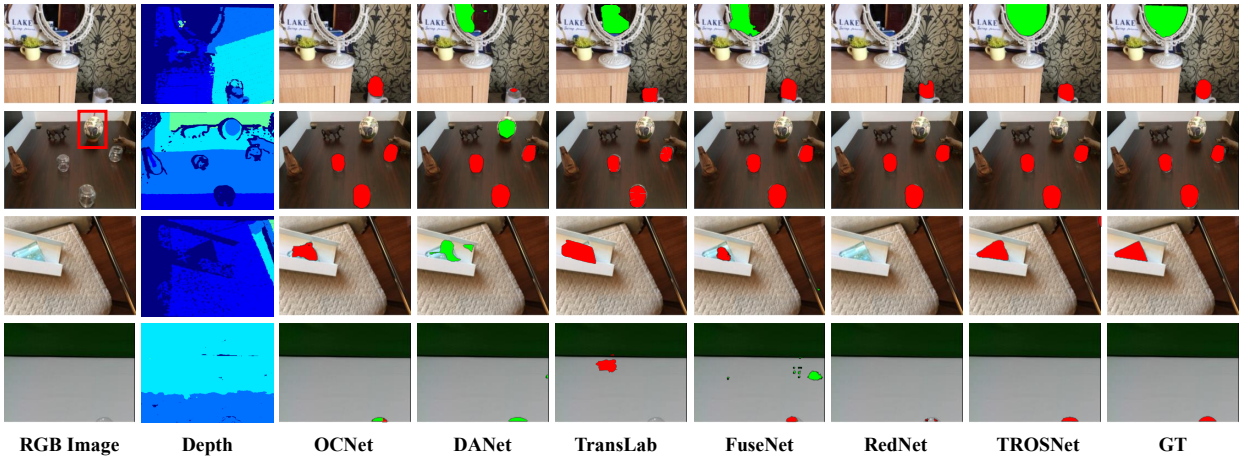

Fig. 13. Visualization of results on TROSD: OCNet [15], DANet [48] and TransLab [8] are RGB based methods; FuseNet [52], RedNet [53] and the proposed TROSNet are RGB-D based methods.

accuracy, TROSNet (81.21%) is about 10.26% higher than the second best method (DANet 70.95%).

Qualitative comparison of these methods are visualized in Fig. 13 on four typical examples. First, it is clear from the 1*st* row of Fig. 13) that only our TROSNet can correctly segment an image containing both transparent and reflective objects. Second, we also show an image including objects that are not transparent or reflective in the scene (the red bounding box in the 2*nd* row). Although DANet [48] shows a promising segmentation result on transparent objects, it suffers from segmenting an object neither transparent nor reflective to be a reflective one. Third, we also consider object occlusion (the 3*rd* row). Although a transparent glass is partly inside a plastic box, our TROSNet can segment its visible part effectively. Fourth, as for some smaller transparent or reflective objects (the 4*th* row), FuseNet and TransLab do not perform well. Compared with other methods, TROSNet still performs well, which can be ascribed to the Feature Fusion module and the BR block.

C. Results on the ClearGrasp dataset

TABLE V
COMPARISON OF TROSNET WITH THE CLEARGRASP METHOD ON THE CLEARGRASP TEST SET WITH FINE-TUNING. KNOWN: KNOWN OBJECTS FROM CLEARGRASP. NOVEL: NOVEL OBJECTS FROM CLEARGRASP.

Method	Input	Training Set	IoU (%) Known	IoU (%) Novel
ClearGrasp [6]	RGB	ClearGrasp	63	58
EBLNet	RGB-D	TROSD	65.01	64.94
TROSNet	RGB-D	TROSD	**71.63**	**71.19**

To validate the generalizability of our baseline method TROSNet for transparent objects only, we train it on the images in the TROSD dataset with only transparent objects and test it on the ClearGrasp test set directly. As shown in Table V, our TROSNet can still achieve a good performance (higher than 71.63% in IoU). What is more, for the novel objects in the ClearGrasp test set, TROSNet (71.19%) performs much better than the ClearGrasp method (58%) in IoU. For qualitative comparison, some typical results are visualized in Fig. 14.

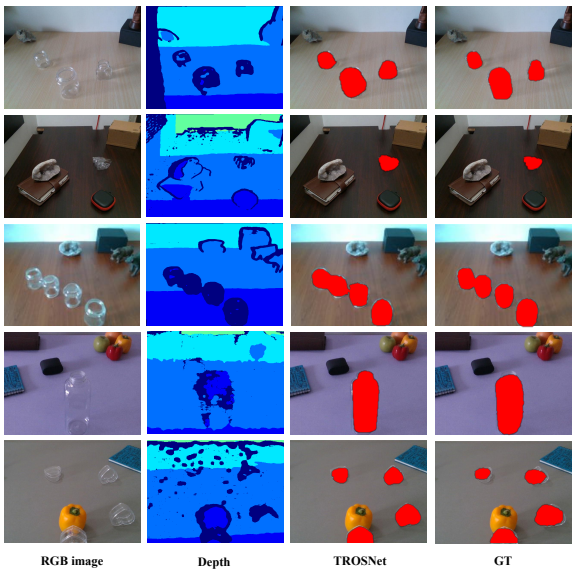

Fig. 14. Visualization of results on the ClearGrasp novel test dataset.

D. Results on a reflective object dataset

To further evaluate the generalizability of TROSNet for reflective objects only, we compare it with MirrorNet [3] and PMD [35], which are RGB based methods and outstanding mirror segmentation networks on the MSD dataset [3].

Firstly, we pick out the images that contain reflective objects (*e.g.*, mirror and metal) from the TROSD dataset. These images form a new training set (3,646 RGB-D images) and a

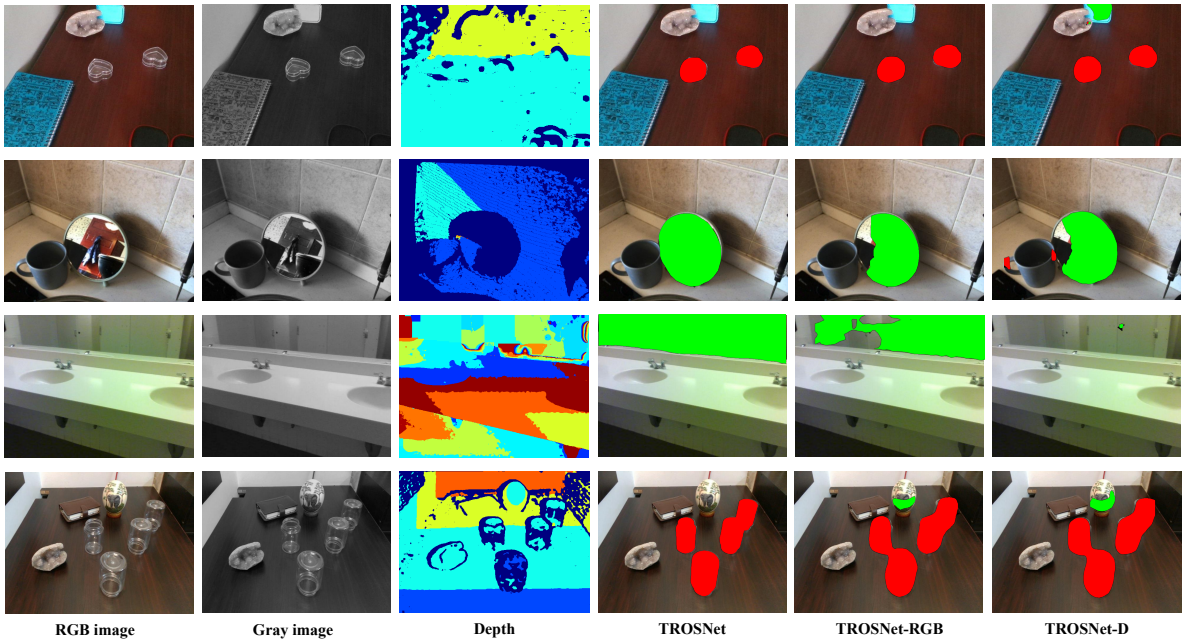

Fig. 15. Visualized results of TROSNet with different inputs.

TABLE VI
COMPARISON OF TROSNET WITH MIRRORNET ON THE REFLECTIVE SUBSET OF TROSD. R: REFLECTIVE OBJECTS. B: BACKGROUND.

Method	Input	IoU (%) R	IoU (%) B	mIoU (%)	mAcc (%)
MirrorNet [3]	RGB	48.28	95.68	71.98	77.55
PMD [35]	RGB	57.18	96.02	76.61	82.37
TROSNet-RGB	RGB	**61.92**	**96.56**	**79.24**	**87.55**

TABLE VII
EFFECTIVENESS OF DIFFERENT INPUT MODALITIES. TROSNET-RGB: WE ONLY FEED RGB IMAGES INTO TROSNET. TROSNET-D: WE ONLY FEED DEPTH IMAGES INTO TROSNET. R: REFLECTIVE OBJECTS. T: TRANSPARENT OBJECTS. B: BACKGROUND.

Method	RGB	Depth	IoU (%) R	IoU (%) T	IoU (%) B	mIoU (%)	mAcc (%)
TROSNet-RGB	✓		48.75	48.56	95.49	64.26	75.93
TROSNet-D		✓	50.15	28.03	94.04	57.41	67.82
TROSNet	✓	✓	**62.27**	**57.23**	**96.52**	**72.01**	**81.21**

new test set (1,332 RGB-D images). This new dataset is similar to the MSD dataset [3]. Afterwards, MirrorNet (an RGB based method) and TROSNet (an RGB-D based method) are trained and tested on the new set. As shown in Table VI, compared with the RGB based methods, our TROSNet achieves much better performance (+13.64% and +4.74% in IoU for reflective objects). This clearly indicates the advantage of our TROSNet-RGB method.

E. Ablation studies

1) Different input modalities: We first conduct ablation studies on the effectiveness of different input modalities to verify the necessity of exploiting both RGB information and depth information. TROSNet-RGB is trained and tested with RGB images as the RGB channel and their corresponding gray images as the depth channel. As for TRSONet-D, we set the input for the RGB channel as all-zero tensor. All experiments in Table VII adopt ResNet-50 [39] as encoder with the same training scheme on TROSD.

As shown in Table VII, TROSNet-RGB (i.e., TROSNet with only RGB images as input) can achieve 75.93% in mAcc, which is higher than that of all other SOTA methods in Table IV (except for our TROSNet).

We can also find that the performance of TROSNet-D (only utilizing depth images as input) on the reflective object segmentation is better than all RGB based SOTA methods, indicating that the depth distortion cues are very helpful for reflective object segmentation.

With both modalities used, TROSNet achieves the best performance of 72.01% in mIoU and 81.21% in mAcc, demonstrating the effectiveness of exploiting both RGB information and depth information for transparent and reflective object segmentation.

Some visualized results are shown in Fig. 15, and the promising performance of TROSNet in the fourth column supports the necessity of utilizing both RGB and depth images. In principle, in RGB-D segmentation, RGB provides semantic information and depth provides spatial information. Absence of either single mode decreases the segmentation performance because of the lack of supplementary information from another

mode. Besides, it might be too difficult to rely on only a noisy mode, such as the depth mode, to locate the accurate position of objects, as shown by the ill performance in Fig. 15.

TABLE VIII
ABLATION STUDIES OF ALL BLOCKS IN TROSNET TESTED ON TROSD. R: REFLECTIVE OBJECTS. T: TRANSPARENT OBJECTS. B: BACKGROUND.

Backbone	CMF	CA	BR	IoU (%) R	IoU (%) T	IoU (%) B	mIoU(%)	mAcc(%)
✓				40.40	51.11	96.51	62.68	69.18
✓	✓			43.70	53.59	96.75	64.68	71.48
✓	✓	✓		54.04	54.50	97.02	68.52	77.30
✓	✓	✓	✓	**62.27**	**57.23**	**96.52**	**72.01**	**81.21**

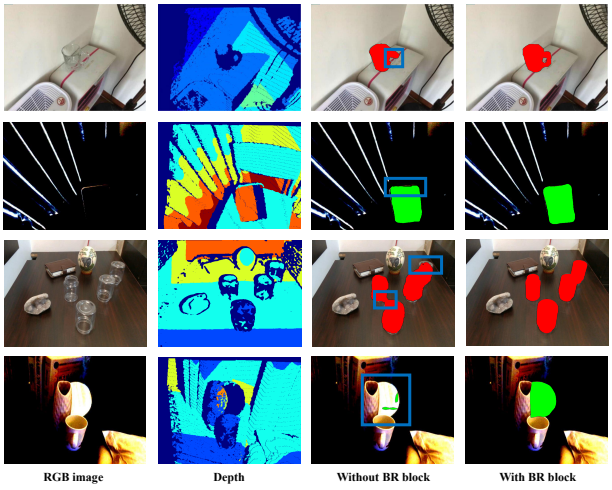

Fig. 16. Effect of the BR block on TROSNet.

2) Component analysis: Then we implement ablation studies on the effects of three blocks (CMF, CA and BR) of TROSNet. We run the models with and without these blocks on TROSD to test how they affect the performance of our TROSNet. We compare the IoU for transparent and reflective objects and the background, as well as the accuracy, for different models. As the results in Table VIII show, the model with all three blocks has the best performance overall, which verifies the necessity of all the three blocks implemented in TROSNet. We also visualize the result w/o BR block in Fig. 16, showing that BR block brings a more refined bound for the output segmentation masks.

TABLE IX
ABLATION STUDIES OF FEATURE FUSION MODULE TESTED ON TROSD. R: REFLECTIVE OBJECTS. T: TRANSPARENT OBJECTS. B: BACKGROUND.

Network Structure	mIoU(%)	mAcc(%)
CEN(with DeepLab v3+)	67.78	73.51
Encoder+CEN+BR+Decoder	69.89	75.62
Encoder+FF+BR+Decoder	**72.01**	**81.21**

3) Additional experiments on feature fusion module: Considering the significance of feature fusion module when facing multi-modal problems [55], we also implement an ablation study on the feature fusion block, which blends the input RGB and depth features and imports the fused features to the next block. We compare our network with CEN, a SOTA feature fusion method [56] which also succeeds in image segmentation. We also adopt this method as a substitute for the feature fusion block in TROSNet, making a comparison of it with our TROSNet on TROSD. The detailed results are recorded in Table IX.

4) Ablation studies on boundary refinement: We also visualize an exemplar result of TROSNet with and without the BR block in Fig. 16. We can see that, besides producing a higher IoU (the network without the BR block fails to recognize the handle as a part of the glass), the BR block helps acquire a smoother boundary in the segmentation result.

TABLE X
ROBUSTNESS STUDIES TROSNET AND EBLNET TESTED ON NOISY TROSD.

Network	RGB w/o noise	Depth w/o noise	IoU (%) R	IoU (%) T	IoU (%) B	mIoU(%)	mAcc(%)
EBLNet	✗	✓	31.76	40.51	92.18	54.82	56.70
EBLNet	✓	✗	46.16	43.58	93.77	61.17	62.39
EBLNet	✗	✗	51.75	50.12	94.57	65.49	67.39
TROSNet	✗	✓	28.33	47.98	95.66	57.82	65.53
TROSNet	✓	✗	50.62	46.37	96.47	64.65	72.29
TROSNet	✗	✗	**62.27**	**57.23**	**96.52**	**72.01**	**81.21**

F. Robustness studies

In collecting real-world RGB-D images, RGB images and depth images are often contaminated by noises, making the downstream operations hard. In this case, we implement a robustness study of our TROSNet and another SOTA segmentation method EBLNet.

We leave the training set unchanged and add Gaussian noise to the test set. The noise is added to RGB images and depth images separately. We test all models under three situations, namely RGB-D without noise, RGB images with noise and depth images with noise.

The results are shown in Table X and Fig. 17. Compared with the SOTA segmentation method EBLNet, our TROSNet possesses a relatively better performance under noisy conditions.

G. Limitations

We also look into some hard scenes for our TROSNet, whose IoU falls below mIoU for entire test set. Some examples are shown in Fig. 18. We notice that most hard scenes come across the situation of occlusion between objects (rows 1 and 2) or appearance of large objects (area ratio of over 30%; rows 3 and 4). Both situations result in serious distortion of depth estimation. Occlusion between objects would trigger distortion due to parallax of RGB-D camera; the masked area of a larger object would contain more noise from depth images, making our model tend to segment the object into separate sub-pixels.

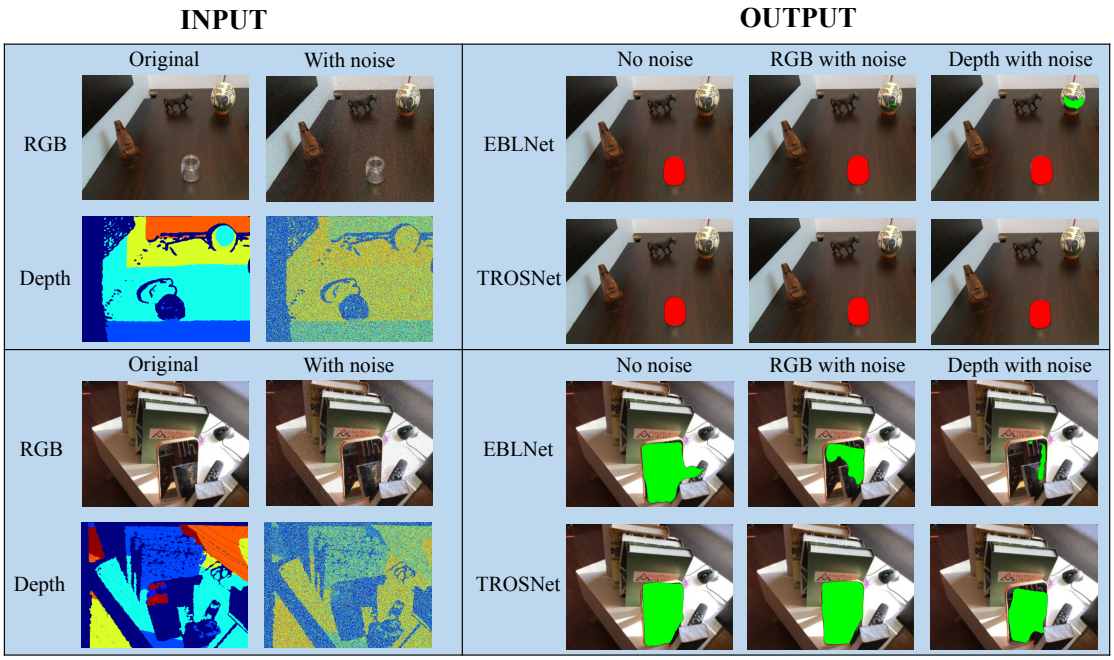

Fig. 17. Results on robustness studies.

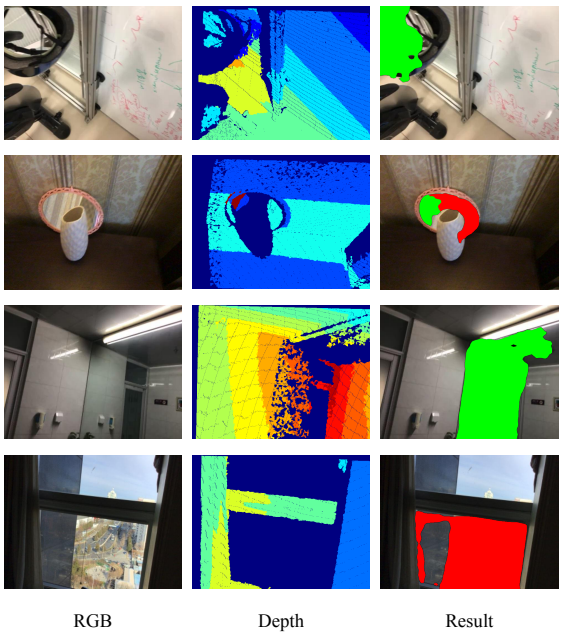

Fig. 18. Results on hard scenes.

VI. Conclusion

In this paper, we introduce TROSD, a new large-scale RGB-D dataset containing 11,060 real-world RGB-D images with detailed annotations for the segmentation of transparent and reflective objects. Along with the dataset, we also propose a high-standard baseline network (TROSNet) with a cascaded multi-modal fusion unit introduced. Extensive experimental results clearly show that TROSNet has the good generalizability to serve as a high-standard baseline method for benchmarking transparent and reflective object segmentation algorithms, and TROSD has the diversity and capacity to serve as a comprehensive test-bed dataset for evaluating and developing new deep networks in this challenging area of transparent and reflective object segmentation in practice. As for future works, to deal with the illness of hard scenes and the existence of noisy depth input, further researches could possibly aim at a feasible solution to alleviate such issues.

References

[1] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[2] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

[3] X. Yang, H. Mei, K. Xu, X. Wei, B. Yin, and R. W. Lau, "Where is my mirror?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8809–8818.

[4] H. Mei, X. Yang, Y. Wang, Y. Liu, S. He, Q. Zhang, X. Wei, and R. W. Lau, "Don't hit me! glass detection in real-world scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3687–3696.

[5] V. Seib, A. Barthen, P. Marohn, and D. Paulus, "Friend or foe: exploiting sensor failures for transparent object localization and classification," in *2016 International Conference on Robotics and Machine Vision*, vol. 10253. SPIE, 2017, pp. 94–98.

[6] S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, "Clear grasp: 3d shape estimation of transparent objects for manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3634–3642.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[8] E. Xie, W. Wang, W. Wang, M. Ding, C. Shen, and P. Luo, "Segmenting transparent objects in the wild," in *European conference on computer vision*. Springer, 2020, pp. 696–711.

[9] A. Kalra, V. Taamazyan, S. K. Rao, K. Venkataraman, R. Raskar, and A. Kadambi, "Deep polarization cues for transparent object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8602–8611.

[10] W. Shi, J. Xu, D. Zhu, G. Zhang, X. Wang, J. Li, and X. Zhang, "RGB-D semantic segmentation and label-oriented voxelgrid fusion for accurate 3D semantic mapping," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 1, pp. 183–197, 2022.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[13] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1925–1934.

[14] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[15] Y. Yuan, L. Huang, J. Guo, C. Zhang, X. Chen, and J. Wang, "Ocnet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916*, 2018.

[16] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 593–602.

[17] M. Everingham, S. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[18] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for rgb-d salient object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*. Springer, 2020, pp. 665–681.

[19] G. Li, Z. Liu, and H. Ling, "Icnet: Information conversion network for rgb-d based salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 4873–4884, 2020.

[20] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3528–3542, 2021.

[21] Y. Yang, Q. Qin, Y. Luo, Y. Liu, Q. Zhang, and J. Han, "Bi-directional progressive guidance network for RGB-D salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5346–5360, 2022.

[22] W. Wang and U. Neumann, "Depth-aware cnn for rgb-d segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 135–150.

[23] J. Jiao, Y. Wei, Z. Jie, H. Shi, R. W. Lau, and T. S. Huang, "Geometry-aware distillation for indoor semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2869–2878.

[24] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and J. Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4106–4115.

[25] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[26] H. Fan, X. Yu, Y. Ding, Y. Yang, and M. Kankanhalli, "Pstnet: Point spatio-temporal convolution on point cloud sequences," *arXiv preprint arXiv:2205.13713*, 2022.

[27] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6411–6420.

[28] F. Li, J. Ma, Z. Tian, J. Ge, H.-N. Liang, Y. Zhang, and T. Wen, "Mirror-yolo: An attention-based instance segmentation and detection model for mirrors," *arXiv preprint arXiv:2202.08498*, 2022.

[29] H. Mei, B. Dong, W. Dong, P. Peers, X. Yang, Q. Zhang, and X. Wei, "Depth-aware mirror segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3044–3053.

[30] "Structure support," *URL: https://structure.io/help/structure-sensor*.

[31] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[32] K. Wada, "Labelme: image polygonal annotation with python. 2016," *URL: https://github.com/wkentaro/labelme*, 2016.

[33] J. Lin, Z. He, and R. W. Lau, "Rich context aggregation with reflection prior for glass surface detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 415–13 424.

[34] J. Lin, Y. H. Yeung, and R. W. Lau, "Depth-aware glass surface detection with cross-modal context mining," *arXiv preprint arXiv:2206.11250*, 2022.

[35] J. Lin, G. Wang, and R. W. Lau, "Progressive mirror detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3697–3705.

[36] J. Tan, W. Lin, A. X. Chang, and M. Savva, "Mirror3d: Depth refinement for mirror surfaces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 990–15 999.

[37] H. Wang, D. Huang, K. Jia, and Y. Wang, "Hierarchical image segmentation ensemble for objectness in RGB-D images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 93–103, 2019.

[38] W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, and W. Lin, "Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 2091–2106, 2022.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[40] H. He, X. Li, G. Cheng, J. Shi, Y. Tong, G. Meng, V. Prinet, and L. Weng, "Enhanced boundary learning for glass-like object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 859–15 868.

[41] Z. Liu, Y. Tan, Q. He, and Y. Xiao, "SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4486–4497, 2022.

[42] J. Kim, M. Kim, H. Kang, and K. H. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *International Conference on Learning Representations*, 2019.

[43] X. Shu, J. Yang, R. Yan, and Y. Song, "Expansion-squeeze-excitation fusion network for elderly activity recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 8, pp. 5281–5292, 2022.

[44] X. Shu, B. Xu, L. Zhang, and J. Tang, "Multi-granularity anchor-contrastive representation learning for semi-supervised skeleton-based action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[45] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659–5667.

[46] J. Zhang, K. Yang, A. Constantinescu, K. Peng, K. Müller, and R. Stiefelhagen, "Trans4Trans: Efficient transformer for transparent object and semantic scene segmentation in real-world navigation assistance," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[47] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[48] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[49] A. Valada, R. Mohan, and W. Burgard, "Self-supervised model adaptation for multimodal semantic segmentation," *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1239–1285, 2020.
[50] W. Zhou, E. Yang, J. Lei, and L. Yu, "Frnet: Feature reconstruction network for rgb-d indoor scene parsing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 4, pp. 677–687, 2022.
[51] D. Seichter, S. Fischedick, M. Köhler, and H.-M. Gross, "Efficient multi-task rgb-d scene analysis for indoor environments," *arXiv preprint arXiv:2207.04526*, 2022.
[52] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture," in *Asian conference on computer vision*. Springer, 2016, pp. 213–228.
[53] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, "Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation," *arXiv preprint arXiv:1806.01054*, 2018.
[54] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
[55] C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2256–2270, 2019.
[56] Y. Wang, F. Sun, W. Huang, F. He, and D. Tao, "Channel exchanging networks for multimodal and multitask dense image prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Tianyu Sun received the B.S. degree in 2021 from the Department of Electronic Engineering, Tsinghua University, where he is currently pursuing the M.S. degree. His research interests include 3D vision and semantic segmentation.

Guodong Zhang received the B.S. degree from the College of Communication Engineering, Jilin University, in 2018. He received the M.S. degree from the Department of Electronic Engineering, Tsinghua University, in 2021. His research interests include deep learning, RGB-D image, and semantic segmentation.

Wenming Yang received his Ph.D. degree in information and communication engineering from Zhejiang University in 2006. He is an Associate Professor in Tsinghua Shenzhen International Graduate School/Department of Electronic Engineering, Tsinghua University. His research interests include image processing, pattern recognition, computer vision and AI in medicine.

Jing-Hao Xue received the Dr.Eng. degree in signal and information processing from Tsinghua University in 1998 and the Ph.D. degree in statistics from the University of Glasgow in 2008. He is a Professor in the Department of Statistical Science at University College London. His research interests include statistical pattern recognition, machine learning and computer vision. He is an Associate Editor of the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Cybernetics, and the IEEE Transactions on Neural Networks and Learning Systems.

Guijin Wang received the B.S. and Ph.D. degrees (with honor) from the Department of Electronic Engineering, Tsinghua University, in 1998 and 2003 respectively, both in signal and information processing. He is a Professor with the Department of Electronic Engineering, Tsinghua University. He was an Associate Editor of IEEE Signal Processing Magazine. His research interests focus on computational imaging, pose recognition, intelligent human-machine UI, intelligent surveillance, industry inspection, and AI for Big medical data.