Power of Boundary and Reflection: Semantic Transparent Object Segmentation using Pyramid Vision Transformer with Transparent Cues

Tuan-Anh Vu[1] Hai Nguyen-Truong[1] Ziqiang Zheng[1] Binh-Son Hua[2]
Qing Guo[3] Ivor W. Tsang[4] Sai-Kit Yeung[1]

[1]Hong Kong University of Science and Technology [2]Trinity College Dublin
[3]Nankai University [4]Centre for Frontier AI Research (CFAR), A*STAR

Abstract

*Glass is a prevalent material among solid objects in everyday life, yet segmentation methods struggle to distinguish it from opaque materials due to its transparency and reflection. While it is known that human perception relies on boundary and reflective-object features to distinguish glass objects, the existing literature has not yet sufficiently captured both properties when handling transparent objects. Hence, we propose incorporating both of these powerful visual cues via the Boundary Feature Enhancement and Reflection Feature Enhancement modules in a mutually beneficial way. Our proposed framework, **TransCues**, is a pyramidal transformer encoder-decoder architecture to segment transparent objects. We empirically show that these two modules can be used together effectively, improving overall performance across various benchmark datasets, including glass object semantic segmentation, mirror object semantic segmentation, and generic segmentation datasets. Our method outperforms the state-of-the-art by a large margin, achieving **+4.2%** mIoU on Trans10K-v2, **+5.6%** mIoU on MSD, **+10.1%** mIoU on RGBD-Mirror, **+13.1%** mIoU on TROSD, and **+8.3%** mIoU on Stanford2D3D, showing the effectiveness of our method against glass objects.*

1. Introduction

Glass objects, such as windows, bottles, walls, and glass, have presented significant challenges to image segmentation due to their appearances being heavily influenced by the surrounding environment. Also, many robotic systems [47, 48, 56] rely on sensor fusion techniques that use sonars or lidars, but these methods often struggle to detect transparent objects and misinterpret reflections as actual objects, leading to scan-matching issues. This is because transparent objects exhibit properties of refraction

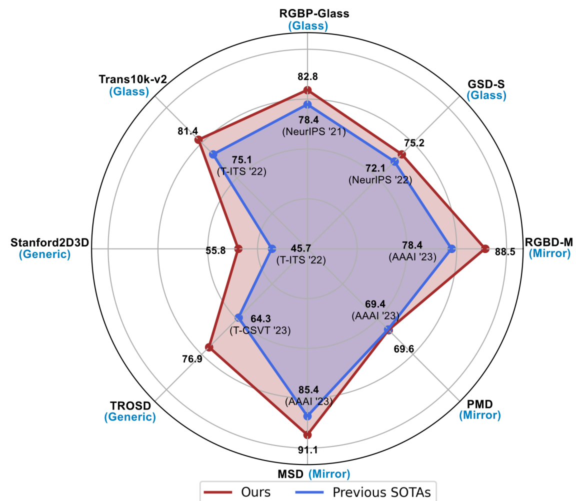

Figure 1. Our method achieves competitive performance compared to previous methods across glass, mirror, and generic segmentation tasks. To maintain fairness, we only compare with methods that use the same input (only RGB image).

and reflection, which cause light to be reflected and to appear in surrounding areas, thereby misleading robot sensors and negatively impacting robot navigation, depth estimation, and 3D reconstruction. Hence, visual systems must deal with reflective surfaces, which would help them accurately identify glass barriers for effective collision prevention in workplaces, supermarkets, or hotels. Furthermore, in domestic and professional settings, visual systems should also be able to navigate fragile items such as vases and glasses. Therefore, a practical, robust, cost-effective, vision-based approach for transparent object segmentation is essential. However, current semantic segmentation algorithms [7, 49, 52, 55, 58] and even powerful foundation models, such as SAM [20] and Semantic SAM [5, 21], were not designed to address transparent and reflective objects, resulting in decreased performance in the presence of such objects. Importantly, the SAM model does not include semantics; in other words, it cannot yield masks with semantic information. The SAM model also presents the challenge of over-segmentation, thereby increasing the likelihood of false positives (see Supplementary for details).

Recent works on the human visual system [19, 34, 35] prove that **"humans rely on specular reflections and boundaries as key indicators of a transparent layer"**. Recent methods for segmenting transparent or glass objects have been proposed, along with various strategies to improve performance when dealing with glass objects. These strategies include focusing on the object's edges [39, 48, 56], using depth information [36, 39], analyzing reflections [24], looking at how light polarizes [29, 46], and utilizing the object's context or semantic information [25]. However, techniques that rely on polarization, depth, and semantic information [25, 29, 36, 39, 46] often require specialized equipment to collect data or extensive human labor to label it, which is inefficient. As far as we know, no method has combined both visual cues of boundary and reflection to improve segmentation performance.

Therefore, we will focus on capturing the two visual cues into the segmentation models: boundary localization for shape inference and reflections for glass surface recognition. We introduce an *efficient transformer-based architecture* tailored for segmenting transparent and reflective objects along with general objects. Then, our method captures the **glass boundaries** based on geometric cues and the **glass reflections** based on appearance cues within an enhanced feature module in our network. In doing so, we developed a Boundary Feature Enhancement (BFE) module to learn and integrate glass-boundary features to improve the localization and segmentation of glass-like regions. We supervise this module with a new boundary loss that uses the Sobel kernel to extract boundaries from the gradients of the predictions and the ground-truth objects' masks. Then, we introduce a Reflection Feature Enhancement (RFE) module that decomposes reflections into foreground and background layers, providing the network with additional features to distinguish glass-like from non-glass areas. By harnessing the power of transformer-based encoders and decoders, our framework can capture long-range contextual information, unlike previous methods that relied heavily on stacked attention layers [12, 51] or on combining CNN backbones with transformers [42, 48, 58]. These long-range visual cues are essential to reliably identify transparent objects, especially when they lack distinctive textures or share similar content with their surroundings [48]. More importantly, we demonstrate that our method is robust to both transparent object segmentation and generic semantic segmentation tasks, with state-of-the-art performance for both scenarios across various datasets.

In summary, our contributions are as follows:

- We introduce **TransCues**, an efficient transformer-based segmentation architecture that segments both transparent, reflective, and general objects.
- We propose the Boundary Feature Enhancement (BFE) module and a boundary loss that improves the accuracy of glass detection.
- We present the Reflection Feature Enhancement (RFE) module, facilitating the differentiation between glass and non-glass regions.
- We conduct extensive experiments to demonstrate our method's competitive performance on diverse tasks, *e.g.* semantic glass segmentation, glass and mirror segmentation, and generic semantic segmentation.

2. Related Works

2.1. Transparent Object Sensing and Segmentation

In the transparency settings, the color intensities of both glass and its background often match, making it challenging to differentiate between them. Traditional visual-aid systems, enhanced with ultrasonic sensors and RGB-D cameras, were developed to effectively identify transparent barriers like glass and windows [18]. On raw images, existing works have explored the use of transmission differences [32], reflection cues [24], and polarization [29, 46] for detecting transparency. Moreover, transparency segmentation methods [4, 6, 15, 30, 33, 48] address a range of objects, from opaque entities (windows and doors) to see-through items (cups and eyeglasses), focusing on discerning reflections and their boundaries to accurately detect and delineate transparent surfaces. Recently, [48] introduced the Trans10K-v2 dataset, which prompts new research directions beyond conventional sensor fusion for transparent objects. This includes AdaptiveASPP [2] for enhanced feature extraction and EBLNet [15] for improved global form representation. Furthermore, Trans4Trans [56] is proposed to provide a lightweight general network for real-world applications. *Building on these innovations*, our work aims to develop an efficient, robust, and transparent object segmentation solution suitable for general semantic segmentation and practical applications like robot navigation.

2.2. Mirror Segmentation

Closely related to glass segmentation is mirror segmentation, in which recent models have introduced high-level concepts to improve detection and localization [13, 16, 17, 40, 45, 50]. For precise localization, SANet [13] utilizes the semantic relationships between mirrors and their surrounding environment. SATNet [17] capitalizes on the natural symmetry between objects and their mirror reflections to accurately identify mirror locations. VCNet [40], in contrast, explores 'visual chirality'—a unique property of mirror images—and incorporates this through a specialized transformation process for effective mirror detection. Lastly, Het-

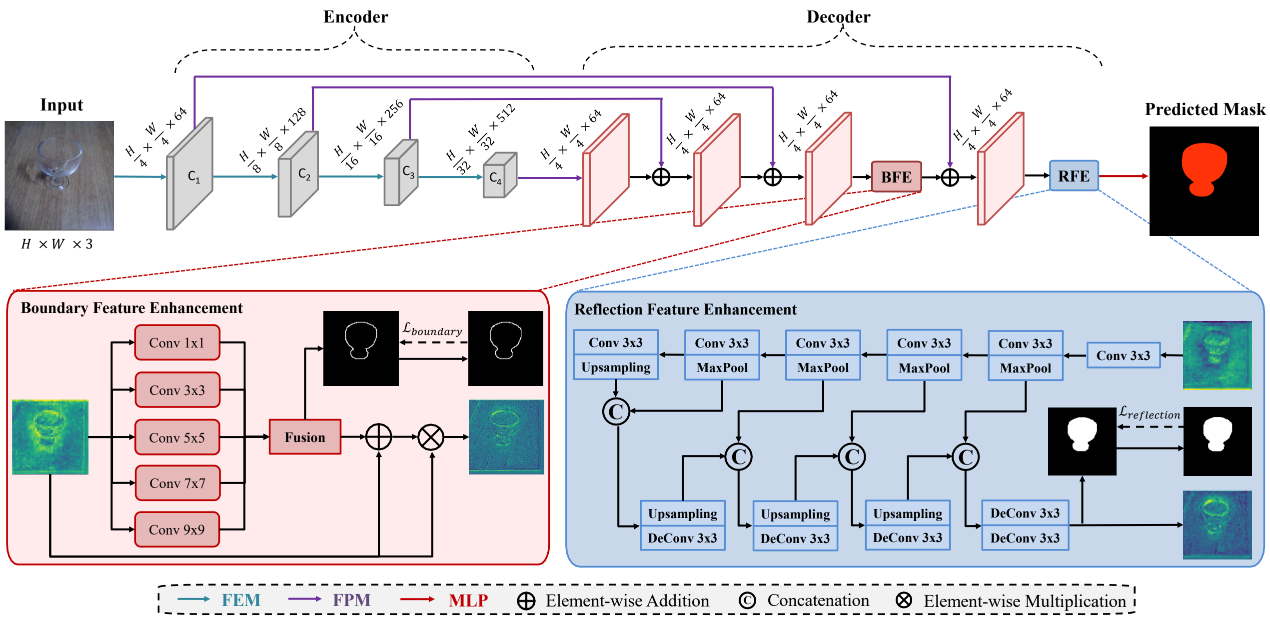

Figure 2. **Overview of our TransCues method.** An RGB image is processed by four FEM modules in the encoder for multi-scale feature extraction. These features are then refined by the decoder's FPM, BFE, and RFE modules, and ultimately converted into semantic labels via an MLP. Our main contributions, BFE and RFE modules, are elaborated in Sections 3.2 and 3.3.

Net [16] introduces a unique model that combines a contrastive module for initial mirror localization and a semantic logical reflection module for semantic analysis. *Unlike existing works*, we consider mirror segmentation as a subproblem in glass segmentation that our proposed framework can also address effectively.

2.3. Transformer in Semantic Segmentation

Since their introduction in natural language processing, transformers have been adopted and further investigated for computer vision tasks. One of the pioneers is Vision Transformer (ViT) [11], which applies transformer layers to sequences of image patches. SETR [58] and Segmenter [38] are inspired by ViT and directly add upsampling and segmentation heads to learn long-range context from the initial layer. MaX-DeepLab [41], and MaskFormer [8] study 2D image segmentation through the perspective of masked prediction and classification based on recent advances of object detection using transformers [3]. As a result, several transformer-based methods for dense image segmentation have been developed [26, 49]. Pyramid architectures for vision transformers have been proposed by [42, 43] to capture hierarchical feature representations. Both ECANet [51] and CSWin transformer [9] recommend applying a self-attention mechanism in either vertical or horizontal stripes to gain advanced simulation capacity while minimizing computing overheads. NAT [14], on the other hand, aims to simplify the standard attention mechanism, resulting in faster processing and reduced memory requirements. Recent methods have been trying to match the performance of transformer models using advanced CNN architectures. MogaNet [22] introduces two feature mixers with depthwise convolutions that efficiently process middle-order information across the spatial and channel spaces. InternImage [44] utilizes deformable convolution, providing a large effective receptive field essential for tasks such as detection and segmentation, and offers adaptive spatial aggregation based on input and task-specifics. *Collectively*, these approaches signify a shift towards more efficient, task-tailored CNN models that strive to replicate the success of transformers in various computer vision applications.

3. Our Proposed Method

Given an RGB image, defined as $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, where H and W respectively denote the image height and width, glass segmentation aims to segment this image into semantic labels at each pixel, which can be expressed as $\mathbf{F} \in \mathbb{R}^{H \times W \times n_{class}}$, where n_{class} represents the number of classes. While glass segmentation can typically be defined in the binary space, defining it along with glass, mirror, and non-glass objects makes for a semantic segmentation problem. Existing work has not considered boundary and reflective cues within the same framework. In particular, while boundary (edge) information was deemed to be captured in [29, 39, 47] for glass segmentation, the reflective information of glass objects was not considered of high priority because of negligible reflections in the datasets. Meanwhile, mirror segmentation has been primarily analyzed for symmetric reflection to detect mirroring in the image [16, 17, 23, 53].

3.1. Our TransCues

As depicted in Figure 2, our network aims to segment glasses in an image and assign their corresponding labels to each pixel. To handle variations in input image sizes across datasets, we standardize the resolution to either 512×512 or 768×768, ensuring consistent position-embedding dimensions throughout training and testing. Our network's architecture adheres to the well-established encoder-decoder structure, comprising the Feature Extraction Module (FEM), Feature Parsing Module (FPM), Boundary Feature Enhancement (BFE), and Reflection Feature Enhancement (RFE) modules. Specifically, the FEM module, based on the PVT architecture [42, 43] in the encoder, efficiently captures multi-scale long-range dependencies in the input image. Drawing inspiration from [56], the FPM module offers a lightweight alternative to the FEM module, capturing detailed-to-abstract representations of transparent objects across C_1 to C_4. The details of FEM and FPM are provided in the Supplementary.

To enrich the feature learning capabilities of our network, we propose capturing the glass boundaries with a *geometric cue* (BFE module) and the glass reflections with an *appearance cue* (RFE module) to differentiate glass from non-glass regions. Boundaries often present as high-contrast edges around transparent objects, a characteristic that aligns well with human visual perception. In detail, we employ the BFE module to amplify the boundary characteristics inherent in transparent features. This enhancement of boundary cues facilitates more accurate segmentation of transparent objects.

Alternatively, reflections on glass surfaces may not always be prominent, making it challenging to design glass segmentation methods. Following this, we feed the boundary-enhanced features into the RFE module. This step is crucial because while most transparent objects exhibit reflections, not all reflective objects are transparent. The RFE module thus plays a pivotal role in distinguishing between these two categories. These cues enhance our network's ability to capture fine details and long-range contextual information for transparent features.

Consequently, as image data progresses through our decoder, it integrates contextual information of varying resolutions, preserving the fine-grained information of transparent and reflective features. By systematically addressing both boundary and reflection cues, our network achieves a nuanced and effective approach to segmenting these challenging object types. Finally, a compact MLP layer is employed to predict semantic labels for each pixel. As shown in Figure 3, throughout each stage of the encoder and the boundary and reflection modules, our feature maps clearly distinguish the glass region, including its defining boundary and reflection, from non-glass surfaces. The following sections will discuss the BFE and RFE modules in more detail.

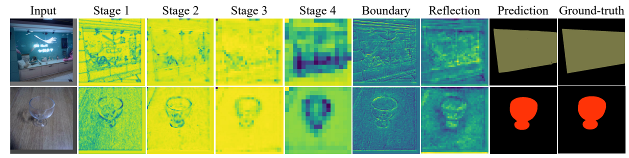

Figure 3. Visualization of feature maps of our method. Zoom in for better visualization.

3.2. Boundary Feature Enhancement Module

Inspired by human perception, incorporating boundary information can significantly benefit segmentation and localization tasks involving glass recognition [15, 30]. To implement this concept, the BFE module is based on the ASPP module [7, 30], yet it is more specialized toward identifying and integrating boundary characteristics of glass into our transformer architecture. Contrary to the approach in [47], which uses an additional boundary stream (an encoder-decoder branch) for boundary feature extraction and integration with primary-stream features, our BFE module is more streamlined. It derives boundary features directly from the targeted input features, bypassing the necessity for an additional stream. As shown in Figure 2, the BFE module is designed to enhance feature learning before the last layer of our decoder so that the reflection module can subsequently improve the features in the next step. We empirically found that this placement of the BFE module achieves better performance and reduced memory usage compared to placing BFE earlier in the decoder (please see Section 4.2 for more details).

The BFE begins by taking input features $\mathcal{X}_0$. These features are then processed through four parallel blocks, each dedicated to extracting multi-scale boundary features $\mathcal{F}_i(.)$ for $i = 1, 2, 3, 4, 5$. Within each block, a convolution layer ($C(.)$) with different kernels and paddings is followed by batch normalization ($BN(.)$) and ReLU activation ($ReLU(.)$) operations, resulting in $\mathcal{F}i = ReLU(BN(C(X)))$. These multi-scale boundary features are subsequently fused using the Fusion module ($\mathcal{F}_{\text{fuse}} = C(\mathcal{F}_1 + \mathcal{F}_2 + \mathcal{F}_3 + \mathcal{F}_4 + \mathcal{F}_5)$), effectively aggregating shape properties and forming the glass boundary features. The output of the Fusion module then undergoes a convolutional layer to predict the boundary map, supervised by the Boundary loss. Finally, the enhanced boundary features $\mathcal{X}_e$ are obtained by aggregating the output of the Fusion module with the input features to locate glass regions, especially their boundaries, as expressed by the following equation:

$$\mathcal{X}_e = (\mathcal{F}_{\text{fuse}}(\mathcal{F}_i(\mathcal{X}_0)) + \mathcal{X}_0) \times \mathcal{X}_0 \tag{1}$$

where $+$ and $\times$ denote element-wise addition and multiplication, respectively.

Boundary loss. The Sobel kernel, also known as the Sobel-Feldman filter, is widely used in image processing and computer vision, primarily for edge detection. It highlights image boundaries by analyzing the 2D gradient and emphasizing high-frequency regions. Our Boundary loss ($\mathcal{L}_b$) leverages the Sobel filter to measure how closely the gradients of a predicted mask match those of the ground truth mask, employing the Dice loss [31]:

$$\mathcal{L}_b = \text{dice}(\nabla_x \hat{M} \oplus \nabla_y \hat{M}, \nabla_x M_{GT} \oplus \nabla_y M_{GT}) \quad (2)$$

where $\hat{M}$ is predicted object mask and M_{GT} is ground truth object mask. ∇_x and ∇_y denote the gradient along x-axis and y-axis computed by the Sobel filter. $\oplus$ represents the combination of the gradient maps into a single feature map. In our implementation, we define the combination $\oplus$ by:

$$a \oplus b = \max\left(\frac{1}{2}(a+b), \tau\right) \quad (3)$$

where τ is set to 0.01 to reduce noise in the gradient maps.

3.3. Reflection Feature Enhancement Module

To enhance the recognition of glass surfaces, we introduce the Reflection Feature Enhancement (RFE) module, which capitalizes on the high reflectivity of glass when illuminated. These reflections provide valuable cues for recognizing glass surfaces in images [28, 53]. Note that if the reflection on the glass surface is insufficient to be discerned by our RFE module, our model may struggle to accurately detect the glass surface. In this scenario, correctly identifying glass surfaces is challenging, even for humans. Please check the Supplementary for discussion and analysis on the need for the RFE module.

In our design, the RFE module is placed after the last layer of the decoder, after the boundary feature enhancement module (please check Section 4.2 for more detail). The RFE module employs a sophisticated convolution-deconvolution architecture [54], which takes input features Y and produces an enhanced feature map Y_e. This architecture allows the module to capture and process information at multiple levels of abstraction, which is essential for handling complex visual cues like reflections. Unlike the other reflection removal models [10, 37, 57] that primarily address global reflections (assuming the entire input image is covered by glass), our RFE module targets detecting local reflections to locate glass surfaces.

In detail, the encoder network $\mathcal{E}$ is responsible for extracting relevant features from the input. It consists of five blocks, each composed of a convolutional layer followed by batch normalization, ReLU activation, and either a Max-Pooling $\mathcal{P}_{max}(.)$ or Upsampling layer $\mathcal{P}_{up}(.)$. Each encoder block can be defined as follows:

$$\mathcal{E}_i = \mathcal{P}^i\left(\text{ReLU}\left(\text{BN}\left(\text{C}(\mathcal{E}_{i-1})\right)\right)\right), \quad i \in [1..5] \quad (4)$$

where $\mathcal{E}_0 = C(Y)$, $\mathcal{P}^i$ is the Max-Pooling or Upsampling layer and when $i = 5$, $\mathcal{P}^i$ will be $\mathcal{P}_{Up}(.)$ instead of $\mathcal{P}_{max}(\cdot)$.

Consequently, the decoder network $\mathcal{D}$ works in conjunction with the encoder to reconstruct and enhance features. It also comprises four blocks, interconnected by an Upsampling layer $\mathcal{P}_{up}(.)$ or a Deconvolutional layer $DC(\cdot)$, along with batch normalization and ReLU activation. Notably, the output of the preceding decoder block and the corresponding feature map $e_i = \mathcal{E}_i$ from the encoder block are concatenated before being fed into the subsequent decoder block. This facilitates the seamless flow of information across the network, enhancing its ability to capture and retain essential features of the reflective areas. Each decoder block can be formulated as follows:

$$\mathcal{D}_j = \mathcal{P}^j(\text{ReLU}(\text{BN}(\text{DC}(\mathcal{D}_{j-1} \frown \mathcal{E}_{5-j})))), \quad j \in [1..4] \quad (5)$$

where $\frown$ is concatentation operation, $\mathcal{D}_0 = \mathcal{E}_5$, $\mathcal{P}^j$ is the Upsampling or Deconvolutional layer and when $j = 4$, $\mathcal{P}^j$ will be $DC(\cdot)$ instead of $\mathcal{P}_{Up}(\cdot)$.

The decoder network's output is split into two tensors: the first tensor represents reflection mask M_{rf}, utilized for optimizing the reflection loss, while the second tensor contains the enhanced reflective features Y_e, which have been processed to capture and emphasize reflection information.

3.4. Loss Functions

We use the softmax cross-entropy loss as our semantic loss $\mathcal{L}_s$ for supervising the semantic mask prediction and the ground truth (GT) semantic mask:

$$\mathcal{L}_s = \text{ce}(\hat{M}, M_{GT}) \quad (6)$$

where $\text{ce}(.)$ is the softmax cross-entropy loss.

To supervise reflection, we also use the softmax cross-entropy loss for our reflection loss $\mathcal{L}_r$. However, there is no GT for the reflection mask, and we assume pseudo GT for the reflection mask to span common categories with reflective appearance, such as window, door, cup, bottle, *etc*. Therefore, we extract pseudo GT with the reflective appearance in the GT semantic map M_{GT}. Note that, as our pseudo GT may contain opaque appearances, we empirically found that this noise is not severe enough to affect the performance of the RFE module. The reflection loss is:

$$\mathcal{L}_r = \text{ce}(M_{\text{rf}}, \phi(M_{GT})) \quad (7)$$

where $\phi(.)$ is a function to extract pseudo GT with reflective appearance in the GT semantic map M_{GT}.
The total loss for our training is:

$$\mathcal{L} = \alpha\mathcal{L}_s + \beta\mathcal{L}_b + \gamma\mathcal{L}_r \quad (8)$$

where α, β and γ are hyper-parameters and are empirically set as [1.0,0.1,0.1] according to the experimental results.

4. Experiments

Datasets. We comprehensively evaluated our proposed method on diverse datasets to demonstrate its exceptional performance and versatility. For details of the datasets, method implementations, experiments, and further analyses, please refer to the supplementary material.

4.1. Qualitative and Quantitative Results

We evaluated our method's performance across three distinct tasks: glass, mirror, and generic segmentation. To ensure **fair comparisons**, we have carefully **selected our model variants** (Ours-X with X is postfixes: -T, -S, -M, -L, -B1, -B2, -B3, -B4, and -B5, represented the model's size as PVTv1 Tiny, Small, Medium, Large, and PVTv2 B1-5, respectively) that have **similar model's size or complexity** used by other methods, as indicated in the respective tables.

Glass Segmentation. We benchmarked our method against recent glass segmentation methods on binary and semantic segmentation tasks. For the binary glass segmentation task, as shown in Table 1, our method (Ours-B4) achieves the pinnacle of mIoU(%) scores, outpacing all other competing methods, which include (SegFormer [49], GSD [24], SETR [58], GDNet [27]). Specifically, it surpasses the runner-up method by margins of 4.35% for RGB-P and 2% for GSD-S. Noteworthy is our method's balance of performance and computational efficiency, which registers relatively lower GFLOPs than its peers. Shifting our focus to the semantic glass segmentation task, where the challenge extends beyond merely detecting glass areas to classifying them into 11 fine-grained categories, our method still reigns supreme. It surpasses competing approaches such as (Trans4Trans [56], DenseASPP [52], DeepLabv3+ [7], OCNet [55], Trans2Seg [48]) by a substantial 4.15% margin in terms of mIoU performance. This dominance in accuracy does not come at the expense of efficiency, as evidenced in Table 2. These comprehensive evaluations underscore the effectiveness of our approach across diverse glass segmentation scenarios, affirming its position as a top-performing and computationally efficient choice for these tasks.

As shown in Figure 4, we observe that recent approaches, such as GDNet and Trans2Seg, may over-detect glass regions in certain images but under-detect them in others, such as GSD. In contrast, our method can accurately identify glass portions of diverse dimensions and morphologies, effectively differentiating them from look-alike non-glass regions in complex images (such as the one showcased in the top right), thanks to the BFE and RFE modules, which leverage boundary and reflection cues, helping our method perform better in challenging scenarios.

Mirror Segmentation. To demonstrate the robustness of our approach for reflective surfaces, we rigorously evaluated our method on three standard binary mirror segmentation datasets: MSD, PMD, and RGBD-Mirror. These datasets are chosen for their resemblance to the reflective characteristics inherent in glass objects. Ours-B3 was chosen as a representative model for a balanced and comparable evaluation. We compared its performance against recent state-of-the-art methodologies, such as SANet [13], VCNet [40], and SATNet [17]. The outcomes, detailed in Table 3 and Figure 5, unequivocally highlight our method's supremacy in mirror segmentation, surpassing the competition across various metrics.

Table 1. Binary Glass Segmentation on RGB-P, GSD-S. We reported mIoU(%) for both datasets.

Method	Backbone	GFLOPs ↓	RGB-P	GSD-S
SegFormer	MiT-B5	70.2	78.4	54.7
Ours-B4	PVTv2-B4	79.3	**82.1**	**74.1**
GSD	ResNeXt-101	92.7	78.1	72.1
SETR	ViT-Large	240.1	77.6	56.7
GDNet	ResNeXt-101	271.5	77.6	52.9

Table 2. Semantic Glass Segmentation on Trans10K-v2.

Method	GFLOPs ↓	Accuracy ↑	mIoU ↑
Trans4Trans-M	34.38	95.01	75.14
DenseASPP	36.20	90.86	63.01
Ours-B2	37.03	**95.92**	**79.29**
DeepLabv3+	37.98	92.75	68.87
OCNet	43.31	92.03	66.31
Trans2Seg	49.03	94.14	72.15

Generic Segmentation. To evaluate our method, we compared it with existing approaches (Trans4Trans [56], TransLab [47], DANet [12], TROSNet [39], Trans2Seg [48], PVT [42]). As detailed in Table 4 and Figure 7, our method outshines SOTA competitors on the TROSD dataset (a dedicated dataset for transparent and reflective object segmentation), underscoring its effectiveness in handling complex transparent and reflective object segmentation. This is mainly because our approach focuses on low-level features, enabling accurate identification and preservation of content differences along the borders of transparent and reflective objects. Moreover, to demonstrate our generalization ability, we evaluate our method on the large-scale real-world Stanford2D3D dataset, which comprises common and transparent objects (less than 1% of the total images). In Table 5, our method outperforms other existing works (about 8.25% better performance in mIOU) in semantic scene segmentation, demonstrating its robustness in discerning specific objects' appearance and normal scenes.

Failure Cases. Figure 6–left shows failure cases of our method and others on Trans10K-v2. Our method would confuse and fail to segment the object with properties similar to those of others. In such a scenario, even humans would struggle to differentiate between these transparent

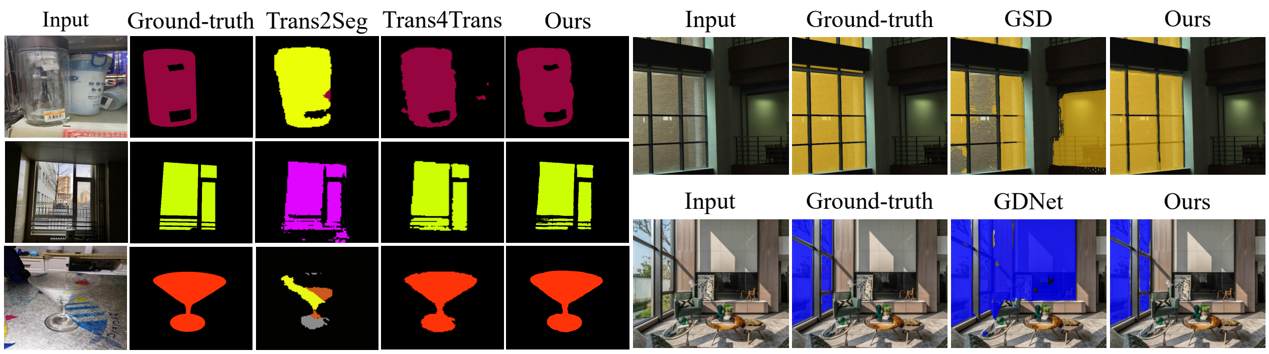

Figure 4. Comparison of glass segmentation methods on Trans10K-v2 (left), RGB-P (top-right), and GSD-S (bottom-right) datasets.

Table 3. Binary Mirror Segmentation on MSD, PMD, and RGBD-Mirror datasets. Our method achieves the best performance in terms of all the evaluation metrics.

Method	Backbone	MSD			PMD			RGBD-Mirror		
		IoU ↑	F_β ↑	MAE ↓	IoU ↑	F_β ↑	MAE ↓	IoU ↑	F_β ↑	MAE ↓
SANet [13]	ResNeXt101	79.85	0.879	0.054	66.84	0.837	0.032	74.99	0.873	0.048
VCNet [40]	ResNeXt101	80.08	0.898	0.044	64.02	0.815	0.028	73.01	0.849	0.052
SATNet [17]	Swin-S	85.41	0.922	0.033	69.38	0.847	0.025	78.42	0.906	0.031
Ours-B3	PVTv2-B3	**91.04**	**0.953**	**0.028**	**69.61**	**0.853**	**0.021**	**88.52**	**0.954**	**0.027**

Figure 5. Qualitative comparison of our method with other methods on MSD, PMD, and RGBD-Mirror datasets.

things. Despite assigning wrong label, our method can still maintain the object's shape. We also show several failure instances (Figure 6–right) in our system that misinterpret non-glass areas as glass because they seem and behave the same, *e.g.* the door frame with reflection and distortion.

Table 4. Comparison of various methods on TROSD dataset. R: reflective objects. T: transparent objects. B: background.

Method	Backbone	IOU ↑ R	IOU ↑ T	IOU ↑ B	mIoU ↑	mAcc ↑
TransLab	ResNet-50	42.57	50.72	96.01	63.11	68.72
DANet	ResNet-101	42.76	54.39	95.88	64.34	70.95
TROSNet	ResNet-50	48.75	48.56	95.49	64.26	75.93
Ours-B3	PVTv2	**67.25**	**67.23**	**97.69**	**77.39**	**87.62**

Table 5. Comparison with state-of-the-art methods on Stanford2D3D dataset.

Method	GFLOPs ↓	MParams ↓	mIoU ↑
Tran4Tran-M	34.38	43.65	45.73
Ours-B2	37.03	27.59	**53.98**
Trans2Seg-M	40.98	30.53	43.83
PVT-M	49.00	56.20	42.49

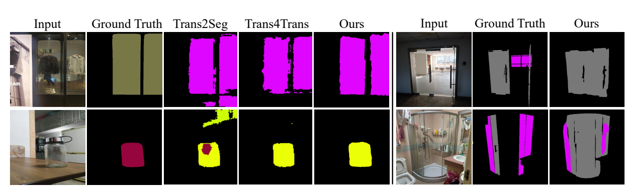

Figure 6. Failure cases of our method and existing methods on the Trans10K-v2 dataset.

4.2. Ablation studies

We present ablation studies to verify various aspects of our model's design and underscore the significance of each module. Any alterations or omissions to the proposed de-

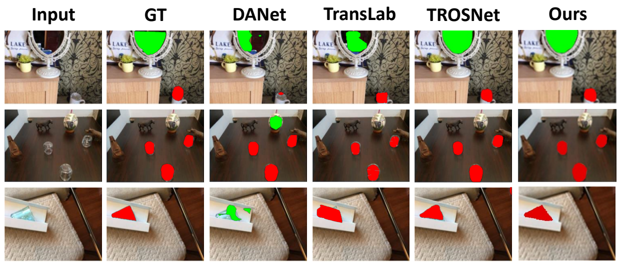

Figure 7. Qualitative comparison of our method with other methods on the TROSD dataset.

sign led to noticeable performance drops, which justifies our choice of the transformer architecture and the boundary and reflection feature learning components.

Effectiveness of different modules. To assess the contribution of both the proposed BFE and RFE modules to our architecture's performance, we systematically evaluated the model under various configurations: ❶ **Baseline Model** (PVTv1-T or PVTv2-B1 without BFE and RFE): this served as our control group, where both the BFE and RFE modules were excluded. Results indicate a foundational performance against which the other configurations could be compared. ❷ **Incorporation of BFE:** When only the BFE module was integrated into our network, we noticed a significant performance enhancement. However, this performance did not reach the potential of the combined BFE and RFE configuration. This proved that while BFE is essential, it is most effective in tandem with RFE. ❸ **Incorporation of RFE:** Similarly, adding only the RFE module to the baseline network also increased performance. This emphasized the value of detecting reflections in transparent objects for the segmentation task. ❹ **Combined Integration of BFE and RFE:** both modules were simultaneously integrated into our network. The performance gain observed in this configuration, as shown in Table 6, was the most pronounced, with gains of **6.36%** and **7.61%** in mIoU on the Trans10K-v2 and Stanford2D3D datasets, respectively. This confirms that the combined effects of boundary and reflection cues significantly augment the network's segmentation capabilities.

Interestingly, the ablation studies further explain why our method performs well for generic segmentation, as demonstrated on the Stanford2D3D dataset. Table 6 shows that the boundary module yields the largest performance gain compared to the reflection module. This means that, for generic segmentation, where reflection feature learning yields negligible improvement, our boundary feature learning remains effective across general semantic labels.

Placement of modules. Incorporating both visual cues into the same framework is nontrivial, as certain features may be best captured at different stages. In our framework, we find the related features better captured at the end of the

Table 6. Effectiveness of different modules of our method on Trans10K-v2 dataset [48] and Stanford2D3D (S2D3D) dataset [1]. We reported mIoU(%) as a metric in this study. The last row corresponds to our method (Ours-B1).

Backbone	FLOPs	Params	BFE	RFE	S2D3D	Trans10K
PVTv1-T	10.16	13.11	-	-	45.19	69.44
PVTv2-B1	11.48	13.89	-	-	46.79 +1.6	70.49 +1.05
PVTv2-B1	13.22	14.37	-	✓	48.12 +2.93	72.65 +3.21
PVTv2-B1	19.55	14.39	✓	-	50.22 +5.03	74.89 +5.45
PVTv2-B1	21.29	14.87	✓	✓	**51.55** +6.36	**77.05** +7.61

Table 7. Placement of modules on Trans10k-v2 dataset. $X \rightarrow Y$ and $X \parallel Y$ means placing X before Y and placing them in parallel and concatenate their outputs in Figure 2.

Variants	MParams ↓	mIoU ↑
Baseline	13.89	70.49
+ RFE → BFE in Encoder	48.98	73.54
+ BFE → RFE in Encoder	48.99	74.12
+ RFE → BFE in Decoder	14.90	75.11
+ BFE ∥ RFE in Decoder	14.91	75.44
TransCues (BFE → RFE in Decoder)	14.87	**77.05**

Decoder layers, with BFE followed by the RFE module. Table 7 shows that the aforementioned modules' positions and processing order matter.

5. Conclusions

In conclusion, this work proposes a method to segment transparent, opaque, and general objects using a pyramidal transformer architecture. Our method exploits two important visual cues, boundary and reflection features, which significantly improve performance in both transparent and generic segmentation tasks. We extensively evaluated our proposed method on several benchmark datasets, demonstrating its robustness in various scenarios. Our architecture is a fully transformer-based method built upon the PVT [42]. Therefore, some limitations remain that reduce our method's capabilities for visual tasks. Firstly, the position encoding in our network is fixed-size, requiring a resizing step that can damage and distort the object's shape. Secondly, similar to other vision transformer-based methods, our network incurs a relatively high computational cost when processing high-resolution images. Finally, as stated earlier, we use the same positional embedding as ViT [11] and PVT [42], which is insufficient for images of arbitrary resolution. In future work, we would like to investigate how to address the above limitations and improve failure cases. In addition, we would like to investigate the extension of our method to other modalities, including depth images, event data, videos, and dynamic scenes.

References

[1] I. Armeni, A. Sax, A. R. Zamir, and S. Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *ArXiv e-prints*, 2017. 8

[2] Yang Cao, Zhengqiang Zhang, Enze Xie, Qibin Hou, Kai Zhao, Xiangui Luo, and Jian Tuo. Fakemix augmentation improves transparent object detection. *arXiv preprint arXiv:2103.13279*, 2021. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[4] Qingling Chang, Huanhao Liao, Xiaofei Meng, Shiting Xu, and Yan Cui. Panoglassnet: Glass detection with panoramic rgb and intensity images. *IEEE Transactions on Instrumentation and Measurement*, 73:1–15, 2024. 2

[5] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. `https://github.com/fudan-zvg/Semantic-Segment-Anything`, 2023. 1

[6] Jiawei Chen, Wen Su, Mengjiao Ge, Ye He, and Jun Yu. To-former: semantic segmentation of transparent object with edge-enhanced transformer. *Vis. Comput.*, 41(3):1811–1825, 2024. 2

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1, 4, 6

[8] Bowen Cheng, A. G. Schwing, and A. Kirillov. Per-pixel classification is not all you need for semantic segmentation. *arXiv*, 2021. 3

[9] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *CVPR*, 2022. 3

[10] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson W.H. Lau. Location-aware single image reflection removal. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4997–5006, 2021. 5

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, Matthias Minderer, Georg H., S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 3, 8

[12] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 2, 6

[13] Huankang Guan, Jiaying Lin, and Rynson W.H. Lau. Learning semantic associations for mirror detection. In *CVPR*, 2022. 2, 6, 7

[14] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *CVPR*, 2023. 3

[15] Hao He, Xiangtai Li, Guangliang Cheng, Jianping Shi, Yunhai Tong, Gaofeng Meng, Véronique Prinet, and LuBin Weng. Enhanced boundary learning for glass-like object segmentation. In *ICCV*, 2021. 2, 4

[16] Ruozhen He, Jiaying Lin, and R.W.H. Lau. Efficient mirror detection via multi-level heterogeneous learning. In *AAAI*, 2023. 2, 3

[17] Tianyu Huang, Bowen Dong, Jiaying Lin, Xiaohui Liu, Rynson W.H. Lau, and Wangmeng Zuo. Symmetry-aware transformer-based mirror detection. In *AAAI*, 2023. 2, 3, 6, 7

[18] Zhiming Huang, Kaiwei Wang, Kailun Yang, Ruiqi Cheng, and Jian Bai. Glass detection and recognition based on the fusion of ultrasonic sensor and RGB-D sensor for the visually impaired. In *Target and Background Signatures IV*. SPIE, 2018. 2

[19] Takahiro Kawabe, Kazushi Maruya, and Shin'ya Nishida. Perceptual transparency from image deformation. *PNAS*, 2015. 2

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1

[21] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity, 2023. 1

[22] Siyuan Li, Zedong Wang, Zicheng Liu, Cheng Tan, Haitao Lin, Di Wu, Zhiyuan Chen, Jiangbin Zheng, and Stan Z. Li. Efficient multi-order gated aggregation network. *ArXiv*, abs/2211.03295v2, 2023. 3

[23] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. Progressive mirror detection. In *CVPR*, 2020. 3

[24] Jiaying Lin, Zebang He, and Rynson W.H. Lau. Rich context aggregation with reflection prior for glass surface detection. In *CVPR*, 2021. 2, 6

[25] Jiaying Lin, Yuen-Hei Yeung, and R.W.H. Lau. Exploiting semantic relations for glass surface detection. *NeurIPS*, 2022. 2

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3

[27] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson W.H. Lau. Don't hit me! glass detection in real-world scenes. In *CVPR*, 2020. 6

[28] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *CVPR*, 2021. 5

[29] Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Glass segmentation using intensity and spectral polarization cues. In *CVPR*, 2022. 2, 3

[30] H. Mei, X. Yang, L. Yu, Q. Zhang, X. Wei, and R. H. Lau. Large-field contextual feature learning for glass detection. *TPAMI*, 2023. 2, 4

[31] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, pages 565–571, 2016. 5

[32] Atsuro Okazawa, Tomoyuki Takahata, and Tatsuya Harada. Simultaneous transparent and non-transparent object segmentation with multispectral scenes. In *IROS*, 2019. 2

[33] Fulin Qi, Xin Tan, Zhizhong Zhang, Mingang Chen, Yuan Xie, and Lizhuang Ma. Glass makes blurs: Learning the visual blurriness for glass surface detection. *IEEE Transactions on Industrial Informatics*, 20(4):6631–6641, 2024. 2

[34] Masataka Sawayama, Yoshinori Dobashi, Makoto Okabe, Kenchi Hosokawa, Takuya Koumura, Toni P. Saarela, Maria Olkkonen, and Shin'ya Nishida. Visual discrimination of optical material properties: A large-scale study. *Journal of Vision*, 22(2):17–17, 2022. 2

[35] Nick Schlüter and Franz Faul. Visual shape perception in the case of transparent objects. *Journal of Vision*, 19(4):24–24, 2019. 2

[36] Daniel Seichter, Söhnke Fischedick, Mona Köhler, and Horst-Michael Gross. Efficient multi-task rgb-d scene analysis for indoor environments. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, 2022. 2

[37] Zhenbo Song, Zhenyuan Zhang, Kaihao Zhang, Wenhan Luo, Zhaoxin Fan, Wenqi Ren, and Jianfeng Lu. Robust single image reflection removal against adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24688–24698, 2023. 5

[38] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, 2021. 3

[39] Tianyu Sun, Guodong Zhang, Wenming Yang, Jing-Hao Xue, and Guijin Wang. TROSD: A new RGB-D dataset for transparent and reflective object segmentation in practice. *IEEE TCSVT*, 2023. 2, 3, 6

[40] Xin Tan, Jiaying Lin, Ke Xu, Pan Chen, Lizhuang Ma, and Rynson W.H. Lau. Mirror detection with the visual chirality cue. *IEEE TPAMI*, 2023. 2, 6, 7

[41] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. MaX-DeepLab: End-to-End panoptic segmentation with mask transformers. In *CVPR*, 2021. 3

[42] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 2, 3, 4, 6, 8

[43] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):1–10, 2022. 3, 4

[44] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, 2023. 3

[45] Yonghui Wang, Wengang Zhou, Yunyao Mao, and Houqiang Li. Detect any shadow: Segment anything for video shadow detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5):3782–3794, 2024. 2

[46] Kaite Xiang, Kailun Yang, and Kaiwei Wang. Polarization-driven semantic segmentation via efficient attention-bridged fusion. *Optica Express*, 2021. 2

[47] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *ECCV*, 2020. 1, 3, 4, 6

[48] Enze Xie, Wenjia Wang, Wenhai Wang, Peize Sun, Hang Xu, Ding Liang, and Ping Luo. Segmenting transparent object in the wild with transformer. In *IJCAI*, 2021. 1, 2, 6, 8

[49] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, José Manuel Álvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 1, 3, 6

[50] Zhifeng Xie, Sen Wang, Qiucheng Yu, Xin Tan, and Yuan Xie. Csfwinformer: Cross-space-frequency window transformer for mirror detection. *IEEE Transactions on Image Processing*, 33:1853–1867, 2024. 2

[51] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelhagen. Capturing omni-range context for omnidirectional segmentation. In *CVPR*, 2021. 2, 3

[52] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. DenseASPP for semantic segmentation in street scenes. In *CVPR*, 2018. 1, 6

[53] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and R.W.H. Lau. Where is my mirror? In *ICCV*, 2019. 3, 5

[54] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 5

[55] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. OCNet: Object context for semantic segmentation. *IJCV*, 2021. 1, 6

[56] Jiaming Zhang, Kailun Yang, Angela Constantinescu, Kunyu Peng, Karin Müller, and Rainer Stiefelhagen. Trans4trans: Efficient transformer for transparent object and semantic scene segmentation in real-world navigation assistance. *IEEE T-ITS*, 2022. 1, 2, 4, 6

[57] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *CVPR*, 2018. 5

[58] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 1, 2, 3, 6