Segmenting Transparent Object in the Wild with Transformer

Enze Xie1*, Wenjia Wang2, Wenhai Wang3, Peize Sun1, Hang Xu1, Ding Liang2, Ping Luo1
1The University of Hong Kong
2Sensetime Retsearch
3Nanjing University
4Huawei Noah’s Ark Lab

Abstract
This work presents a new fine-grained transparent object segmentation dataset, termed Trans10K-v2, extending Trans10K-v1, the first large-scale transparent object segmentation dataset. Unlike Trans10K-v1 that only has two limited categories, our new dataset has several appealing benefits. (1) It has 11 fine-grained categories of transparent objects, commonly occurring in the human domestic environment, making it more practical for real-world application. (2) Trans10K-v2 brings more challenges for the current advanced segmentation methods than its former version. Furthermore, a novel Transformer-based segmentation pipeline termed Trans2Seg is proposed. Firstly, the Transformer encoder of Trans2Seg provides the global receptive field in contrast to CNN’s local receptive field, which shows excellent advantages over pure CNN architectures. Secondly, by formulating semantic segmentation as a problem of dictionary look-up, we design a set of learnable prototypes as the query of Trans2Seg’s Transformer decoder, where each prototype learns the statistics of one category in the whole dataset. We benchmark more than 20 recent semantic segmentation methods, demonstrating that Trans2Seg significantly outperforms all the CNN-based methods, showing the proposed algorithm’s potential ability to solve transparent object segmentation. Code is available in github.com/xieenze/Trans2Seg.

1 Introduction
Modern robots, mainly mobile robots and mechanical manipulators, would benefit a lot from the efficient perception of the transparent objects in residential environments since the environments vary drastically. The increasing utilization of glass wall and transparent door in the building interior and the glass cups and bottles in residential rooms has resulted in the wrong detection in various range sensors. In robotic research, most systems perceive the environment by multi-data sensor fusion via sonars or lidars. The sensors are relatively consistent in detecting opaque objects but are still affected by the scan mismatching due to transparent objects. The unique feature of reflection, refraction, and light projection from the transparent objects may confuse the sensors. Thus a reliable vision-based method, which is much cheaper and more robust than high-precision sensors, would be efficient.

Although some transparent objects dataset [Chen et al., 2018a; Mei et al., 2020] were proposed, there are some obvious problems. (1) Limited dataset scale. These datasets often have less than 1K images captured from the real-world and less than 10 unique objects. (2) Poor diversity. The scene of these datasets is monotonous. (3) Fewer classes. All these datasets have only two classes, background and transparent objects. They lack fine-grained categories, which limited their practicality. Recently, [Xie et al., 2020] proposed
a large-scale and high-diversity dataset termed Trans10K, which divide transparent objects as ‘Things’ and ‘Stuff’. The dataset is high diversity, but it also lacks fine-grained transparent categories.

In this paper, we propose a fine-grained transparent object segmentation dataset termed Trans10K-v2 with more elaborately defined categories. The images are inherit from Trans10K-v1 [Xie et al., 2020]. We annotate the 10428 images with 11 fine-grained categories: shelf, jar, freezer, window, glass door, eyelash, cup, glass wall, glass bowl, water bottle, storage box. In Trans10K-v1, transparent things are defined to be grabbed by the manipulators and stuff are for robot navigation. Though two basic categories can partially help robots to interact with transparent objects, the provided fine-grained classes in Trans10K-v2 can provide more. We analyze these objects’ functions and how robots interact with them in appendix.

Based on this challenging dataset, we design Trans2Seg, introducing Transformer into segmentation pipeline for its encoder-decoder architecture. First, the Transformer encoder provides a global receptive field via self-attention. Larger receptive field is essential for segmenting transparent objects because transparent objects often share similar textures and context with its surroundings. Second, the decoder stacks successive layers to interact query embedding with Transformer encoder output. To facilitate the robustness of transparent objects, we carefully design a set of learnable class prototype embeddings as the query for Transformer decoder, and the key is the feature map from the Transformer encoder. Compared with convolutional paradigm, where the class prototypes is the fixed parameters of convolution kernel weight, our design provides a dynamic and context-aware implementation. As shown in Figure 1b, we train and evaluate 20 existing representative segmentation methods on Trans10K-v2, and found that simply applying previous methods to this task is far from sufficient. By successfully introducing Transformer into this task, our Trans2Seg significantly surpasses the best TransLab [Xie et al., 2020] by a large margin (72.1 vs. 69.0 on mIoU).

In summary, our main contributions are three-folds:

- We propose the largest glass segmentation dataset (Trans10K-v2) with 11 fine-grained glass image categories with diverse scenarios and high resolution. All the images are elaborately annotated with fine-shaped masks and function-oriented categories.
- We introduce a new Transformer-based network for transparent object segmentation with Transformer encoder-decoder architecture. Our method provides a global receptive field and is more dynamic in mask prediction, which shows excellent advantages.
- We evaluate more than 20 semantic segmentation methods on Trans10K-v2, and our Trans2Seg significantly outperforms these methods. Moreover, we show this task largely unsolved. Thus more research is needed.

2 Related Work

Semantic Segmentation. In deep learning era, convolutional neural network (CNN) puts forwards the development of semantic segmentation in various datasets, such as ADE20K, CityScapes and PASCAL VOC. One of the pioneer works approaches, FCN [Long et al., 2015], transfers semantic segmentation into an end-to-end fully convolutional classification network. For improving the performance, especially around object boundaries, [Chen et al., 2017; Lin et al., 2016] propose to use structured prediction module, conditional random fields (CRFs) [Chen et al., 2014], to refine network output. Dramatic improvements in performance and inference speed have been driven by aggregating features at multiples scales, for example, PSPNet [Zhao et al., 2017] and DeepLab [Chen et al., 2017; Chen et al., 2018b], and propagating structured information across intermediate CNN representations [Gadde et al., 2016; Liu et al., 2017; Wang et al., 2018].

Transparent Object Datasets. [Chen et al., 2018a] proposed TOM-Net. It contains 876 real images and 178K synthetic images which are generated by POV-Ray. However, only 4 unique objects are used in synthesizing the training data. Recently, [Xie et al., 2020] introduce the first large-scale real-world transparent object segmentation dataset, termed Trans10K. It has 10K+ images while with only 2 categories. In this work, our Trans10K-v2 inherited the data and annotates 11 fine-grained categories.

Transformer in Vision Tasks. Transformer [Vaswani et al., 2017] has been successfully applied in both high-level vision and low-level vision [Han et al., 2020]. In ViT [Dosovitskiy et al., 2020], Transformer is directly applied to sequences of image patches to complete image classification. In object detection areas [Carion et al., 2020; Zhu et al., 2020], DETR reasons about the relations of the object queries and the global image context via Transformer and outputs the final set of predictions in parallel without non-maximum suppression (NMS) procedures and anchor generation. SETR [Zheng et al., 2020] views semantic segmentation from a sequence-to-sequence perspective with Transformer. IPT [Chen et al., 2020] applies Transformer model to low-level computer vision task, such as denoising, super-resolution and deraining. In video processing, Transformer has received significantly growing attention. VisTR [Wang et al., 2020] accomplishes instance sequence segmentation by Transformer. Multiple-object tracking [Sun et al., 2020; Meinhardt et al., 2021] employs Transformers to decode object queries and feature queries of the previous frame into bounding boxes of the current frame, and merged by Hungarian Algorithm or NMS.

3 Trans10K-v2 Dataset

Dataset Introduction. Our Trans10K-v2 dataset is based on Trans10K dataset [Xie et al., 2020]. Following Trans10K, we use 5000, 1000 and 4428 images in training, validation and testing respectively. The distribution of the images is abundant in occlusion, spatial scales, perspective distortion. We further annotate the images with more fine-grained categories due to the functional usages of different objects. Trans10K-v2 dataset contains 10,428 images, with two main categories and 11 fine-grained categories: (1)
Transparent **Things** contain cup, bottle, jar, bowl and eyeglass. (2) **Transparent Stuff** contain windows, shelf, box, freezer, glass walls and glass doors. In respect to fine-grained categories and high diversity, Trans10K-v2 is very challenging, and have promising potential in both computer vision and robotic researches.

**Annotation Principle.** The transparent objects are manually labeled by expert annotators with professional labeling tool. The annotators were asked to provide more than 100 points when they trace the boundaries of each transparent object, which ensures the high-quality outline of the mask shapes. The way of annotation is mostly the same with semantic segmentation datasets such as ADE20K. We set the background with 0, and the 11 categories from 1 to 11. We also provide the scene environment of each image locates at. The annotators are asked to strictly following principles when they label the images: (I) Highly transparent pixels of objects no matter made of glass, plastics or crystals are annotated as masks, other semi-transparent and non-transparent pixels are ignored. (II) When occluded by opaque objects, the pixels will be cropped from the masks. (III) The detailed principle of how we categorize the objects is listed in appendix.

**Dataset Statistics.** The statistic information of CMCC, imaga number, pixel proportion are listed in Table 1 in detail. From Table1, the sum of all the image numbers is larger than 10428 since some image has multiple categories of objects. See the caption for detail.

**Evaluation Metrics.** Results are reported in three metrics that are widely used in semantic segmentation to benchmark the performance of fine-grained transparent object segmentation. (1) **Pixel Accuracy** indicates the proportion of correctly classified pixels. (2) **Mean IoU** indicates mean intersection over union. (3) **Category IoU** indicates the intersection over union of each category.

### 4 Method

#### 4.1 Overall Pipeline

The overall Trans2Seg architecture contains a CNN backbone, an encoder-decoder Transformer, and a small convolutional head, as shown in Figure 3. For an input image of \((H, W, 3)\),

- The CNN backbone generates image feature map of \((\frac{H}{16}, \frac{W}{16}, C)\).
- The encoder takes in the summation of flattened feature of \((\frac{H}{16}, \frac{W}{16}, C)\) and positional embedding of \((\frac{H}{16}, \frac{W}{16}, C)\), and outputs encoded feature of \((\frac{H}{16}, \frac{W}{16}, C)\).
- The decoder interacts the learned class prototypes of \((N, C)\) with encoded feature, and generates attention map of \((N, M, \frac{H}{16}, \frac{W}{16})\), where \(N\) is number of categories, \(M\) is number of heads in multi-head attention.
- The small convolutional head up-samples the attention map to \((N, M, \frac{H}{4}, \frac{W}{4})\), fuses it with high-resolution feature map Res2 and outputs attention map of \((N, \frac{H}{4}, \frac{W}{4})\).

The final segmentation is obtained by pixel-wise argmax operation on the output attention map.
Class Prototypes

channels
predicted mask
3 x H x W

F
this figure, the segmentation mask of two categories (small conv head refer to Figure 4. Finally, we can get the predict results by doing pixel-wise argmax on the attention map. For example, in this figure, the segmentation mask of two categories (Bottle and Eyeglass) corresponds to two class prototypes with same colors.

Figure 3: The whole pipeline of our hybrid CNN-Transformer architecture. First, the input image is fed to CNN to extract features F. Second, for Transformer encoder, the features and position embedding are flatten and fed to Transformer for self-attention, and output feature(Fcls) from Transformer encoder. Third, for Transformer decoder, we specifically define a set of learnable class prototype embeddings(Ecls) as query, Fe as key, and calculate the attention map with Ecls and Fe. Each class prototype embedding corresponds to a category of final prediction. We also add a small conv head to fuse attention map and Res2 feature from CNN backbone. Details of Transformer decoder and small conv head refer to Figure 4. Finally, we can get the predict results by doing pixel-wise argmax on the attention map. For example, in this figure, the segmentation mask of two categories (Bottle and Eyeglass) corresponds to two class prototypes with same colors.

4.2 Encoder

The Transformer encoder takes a sequence as input, so the spatial dimensions of the feature map (H/W/C) is flattened into one dimension(H/W/C). To compensate missing spatial dimensions, positional embedding [Gehring et al., 2017] is supplemented to one dimension feature to provide information about the relative or absolute position of the feature in the sequence. The positional embedding has the same dimension (H/W/C) with the flattened feature. The encoder is composed of stacked encoder layers, each of which consists of a multi-head self-attention module and a feed forward network [Vaswani et al., 2017].

4.3 Decoder

The Transformer decoder takes input a set of learnable class prototype embeddings as query, denoted by Ecls, the encoded feature as key and value, denoted by Fe, and output the attention map followed by Small Conv Head to obtain final segmentation result, as shown in Figure 4.

The class prototype embeddings are learned category prototypes, updated iteratively by a series of decoder layers through multi-head attention mechanisms. We denoted iterative update rule by ⊙, then the class prototype in each decoder layer is:

$$E_{cls}^e = \bigoplus_{i=0,\ldots,s-1} \text{softmax}(E_{cls}^{-1})(E_{cls}^{i}F_{e})$$ (1)

In the final decoder layer, the attention map is extracted out to into small conv head:

$$\text{attention map} = E_{cls}^{e}F_{e}$$ (2)

The pseudo code of small conv head is shown in shown in Figure 4. The attention map from Transformer decode is the shape of (N, M, H/W), where N is number of categories, M is number of heads in multi-head attention. It is up-sampled to (N, M+1, H/W), then fused with high-resolution feature map Res2 in the second dimension to (N, M+1, H/W), and finally transformed into output attention map of (N, H/W).

4.4 Discussion

The most related work with Trans2Seg is SETR and DETR [Zheng et al., 2020; Carion et al., 2020]. In this section we discuss the relations and differences in details.
Trans2Seg and SETR are both segmentation pipelines. Their key difference is reflected in the design of the decoder. In SETR, the decoder is simple several convolutional layers, which is similar with most previous methods. However, the decoder of Trans2Seg is also Transformer, which fully utilizes the advantages of attention mechanism in semantic segmentation.

DETR. Trans2Seg and DETR share similar components in the pipeline, including CNN backbone, Transformer encoder and decoder. The biggest difference is the definition of query. In DETR, the decoder’s queries represents N learnable objects because DETR is designed for object detection. However, in Trans2Seg, the queries represents N learnable class prototypes, where each query represents one category. We could see that the minor change on query design could generalize Transformer architecture to apply to diverse vision tasks, such as object detection and semantic segmentation.

5 Experiments

5.1 Implementation Details.

We implement Trans2Seg with Pytorch. The ResNet-50 [He et al., 2016] with dilation convolution at last stage is adopted as the CNN extractor. For loss optimization, we use Adam optimizer with epsilon 1e-8 and weight decay 1e-4. Batch size is 8 per GPU. We set learning rate 1e-4 and decay by the poly strategy [Yu et al., 2018] for 50 epochs. We use 8 V100 GPUs for all experiments. For all CNN based methods, we random scale and crop the image to 480 × 480 in training, and resize image to 513 × 513 in inference, following common setting on PASCAL VOC [Everingham and Winn, 2011]. For our Trans2Seg, we adopt Transformer architecture and need to keep the shape of learned position embedding same in training/inference, so we directly resize the image to 512 × 512. Code has been released for community to follow.

5.2 Ablation Studies.

We use the FCN [Long et al., 2015] as our baseline. FCN is a fully convolutional network with very simple design, and it is also a very classic semantic segmentation method. First, we demonstrate that Transformer encoder can build long range attention between pixels, which has much larger receptive field than CNN filters. Second, we remove the CNN decoder in FCN and replace by our Transformer decoder, we design a set of learnable class prototypes as queries and show that this design further helps improve the accuracy. Third, we verify our method with Transformer at different scales.

Self-Attention of Transformer Encoder. As shown in Table 2, the FCN baseline without Transformer encoder achieves 62.7% mIoU. When adding Transformer encoder, the mIoU directly improves 6.1%, achieving 66.8% mIoU. It demonstrates that the self-attention module in Transformer encoder provides global receptive filed, which is better than CNN’s local receptive field in transparent object segmentation.

Category Prototypes of Transformer Decoder. In Table 2, we verify the effectiveness of learnable category prototypes in Transformer decoder. In row 2, with traditional CNN decoder, the mIoU is 68.8%. However, with our Transformer decoder, the mIoU boosts up to 72.1% with 3.3% improvement. The strong performance benefits from the flexible representation that learnable category prototypes as queries to find corresponding pixels in feature map.

Scale of Transformer. The scale of Transformer is mainly influenced by three hyper-parameters: (1) Embedding dim of feature. (2) Number of attention layers. (3) MLP ratio in feed forward layer. We are interested in whether enlarging the model size can continuously improve performance. So we set three combinations, as shown in Figure 3. We can find that with the increase of the size of Transformer, the mIoU first increases then decreases. We argue that without massive pre-trained data, e.g. the large-scale nlp data BERT [Devlin et al., 2019] used, the size of Transformer is not the larger the better for our task.

5.3 Comparison to the state-of-the-art.

We select more than 20 semantic segmentation methods: Translab, Deeplabv3+, DABNet, PSPNet, OCNet, DenseA- spp, FCN, UNet, BiseNet, RefineNet, HardNet, HRNet, FastSCNN, ContextNet, LedNet, DUNet, ICNet, FPENet, DFANet, DANet, ESPNetv2, to evaluate on our Trans10K-v2 dataset, the methods selection largely follows the benchmark of TransLab. For fair comparison, we train all the methods with 50 epochs.

Table 4 reports the overall quantitative comparison results on test set. Our Trans2Seg achieves state-of-the-art 72.15% mIoU and 94.14% pixel ACC, significant outperforms other pure CNN-based methods. For example, our method is 2.1% higher than TransLab, which is the previous SOTA method. We also find that our method tend to performs much better on small objects, such as ‘bottle’ and ‘eyeglass’ (10.0% and 5.0% higher than previous SOTA). We consider that the Transformer’s long range attention benefits the small transparent object segmentation.

In Figure 5, we visualize the mask prediction of Trans2Seg and other CNN-based methods. We can find that owing to Transformer’s large receptive field and attention mechanism, our method can distinguish background and different categories transparent objects much better than other methods.
<table>
<thead>
<tr>
<th>Method</th>
<th>FLOPs</th>
<th>ACC↑ mIoU↑</th>
<th>Category IoU↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPENet</td>
<td>0.76</td>
<td>70.31</td>
<td>10.14</td>
</tr>
<tr>
<td>ESPNetv2</td>
<td>0.83</td>
<td>73.03</td>
<td>12.77</td>
</tr>
<tr>
<td>ContextNet</td>
<td>0.87</td>
<td>86.75</td>
<td>46.69</td>
</tr>
<tr>
<td>FastSCNN</td>
<td>1.01</td>
<td>88.05</td>
<td>51.93</td>
</tr>
<tr>
<td>DFNNet</td>
<td>1.02</td>
<td>85.15</td>
<td>42.54</td>
</tr>
<tr>
<td>ENet</td>
<td>2.09</td>
<td>71.67</td>
<td>22.77</td>
</tr>
<tr>
<td>HRNet_w18</td>
<td>4.20</td>
<td>89.58</td>
<td>54.25</td>
</tr>
<tr>
<td>HardNet</td>
<td>4.42</td>
<td>90.19</td>
<td>56.19</td>
</tr>
<tr>
<td>DABNet</td>
<td>5.18</td>
<td>77.43</td>
<td>15.27</td>
</tr>
<tr>
<td>LEDNet</td>
<td>6.23</td>
<td>86.07</td>
<td>46.40</td>
</tr>
<tr>
<td>ICNet</td>
<td>10.64</td>
<td>78.23</td>
<td>23.39</td>
</tr>
<tr>
<td>BiSeNet</td>
<td>19.91</td>
<td>89.13</td>
<td>58.40</td>
</tr>
<tr>
<td>DenseASPP</td>
<td>36.20</td>
<td>90.86</td>
<td>63.01</td>
</tr>
<tr>
<td>DeepLabv3+</td>
<td>37.98</td>
<td>92.75</td>
<td>68.87</td>
</tr>
<tr>
<td>FCN</td>
<td>42.23</td>
<td>91.65</td>
<td>62.75</td>
</tr>
<tr>
<td>OCNNet</td>
<td>43.31</td>
<td>92.03</td>
<td>66.31</td>
</tr>
<tr>
<td>RefineNet</td>
<td>44.56</td>
<td>87.89</td>
<td>58.18</td>
</tr>
<tr>
<td>Translab</td>
<td>61.31</td>
<td>92.67</td>
<td>69.00</td>
</tr>
<tr>
<td>DUNet</td>
<td>123.69</td>
<td>90.67</td>
<td>59.01</td>
</tr>
<tr>
<td>UNet</td>
<td>124.55</td>
<td>82.20</td>
<td>29.23</td>
</tr>
<tr>
<td>DAnet</td>
<td>198.00</td>
<td>92.70</td>
<td>68.81</td>
</tr>
<tr>
<td>PSPNet</td>
<td>187.03</td>
<td>92.37</td>
<td>68.23</td>
</tr>
</tbody>
</table>

Trans2Seg 49.03 94.14 72.15

Table 4: Evaluated state-of-the-art semantic segmentation methods. Sorted by FLOPs. Our proposes Trans2Seg surpasses all the other methods in pixel accuracy and mean IoU, as well as most of the category IoUs (8 in 11).

especially when the image contains multiple objects of different categories. Moreover, our method can obtain high quality detail information, e.g., boundary of object, and tiny transparent objects, while other CNN-based methods fail to do so. More results are shown in supplementary material.

6 Conclusion

In this paper, we present a challenging fine-grained transparent object segmentation dataset with 11 common categories, termed Trans10K-v2. Moreover, we propose a Transformer-based pipeline with encoder having global receptive field and decoder with category query, termed Trans2Seg, to solve this challenging task. Finally, we evaluate more than 20 mainstream semantic segmentation methods and shows that our Trans2Seg clearly surpass these CNN-based segmentation methods.

In the future, we are interested in exploring Transformer encoder-decoder on general segmentation tasks, such as Cityscapes and PASCAL VOC. We will also put more efforts to solve transparent object segmentation task.

Acknowledgements

This work is partially supported by the General Research Fund of Hong Kong No. 27208720, and the Research Donation from Huawei. We Thank Yaqun Liu for insightful discussion.
References


[Chen et al., 2014] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. 2014.


