Supplementary Material for One-stage Progressive Dichotomous Segmentation

Jing Zhu
jingzhu@nyu.edu
Karim Ahmed
k.ahmed1@samsung.com
Wenbo Li
wenbo.li1@samsung.com
Yilin Shen
yilin.shen@samsung.com
Hongxia Jin
hongxia.jin@samsung.com

Samsung Research America

1 Framework of Feature Extractor

Feature extractor plays a crucial role in our model as it provides effective multi-scale features to support dichotomous segmentation. It is also lightweight enough to directly handle high-resolution inputs. To achieve this, we introduce the multi-scale convolutional attention (MS Conv. Attn.) [1] into our feature extractor, which efficiently extracts multi-scale features with lower computation complexity compared to a transformer. The detailed framework of our feature extractor is depicted in Fig. 1.

Initially, a high-resolution image is passed through two convolutional layers with a kernel size of 3 and a stride size of 2, resulting in a 64-dimensional feature map. This 64-dimensional feature map undergoes four levels of multi-scale convolutional attention to generate a 64-dimensional (F_1 in the paper), a 128-dimensional (F_2 in the paper), a 320-dimensional (F_3 in the paper) and a 512-dimensional (F_4 in the paper) feature map, respectively. Each multi-scale convolutional attention level includes a convolutional layer and multiple multi-scale convolutional attention (MS Conv. Attn.) blocks. The kernel size of the convolutional layer is set to 7 for the first level and 3 for the subsequent levels, while the stride size is 4 for the first level and 2 for the remaining levels.

The structure of the multi-scale convolutional attention (MS Conv. Attn.) block is illustrated by the middle blue block in Fig. 1 with each block containing a multi-scale convolutional attention (MS Conv. Attn.) module (shown as the right block of the figure). The multi-scale convolutional attention module consists of seven depth-wise convolutional layers with kernel size of [5x5, 1x7, 1x11, 1x21, 7x1, 11x1, 21x1], along with a multilayer perceptron (MLP) layer. The number of output channels in each layer within the multi-scale convolutional attention (MS Conv. Attn.) block, including those within the multi-scale convolutional

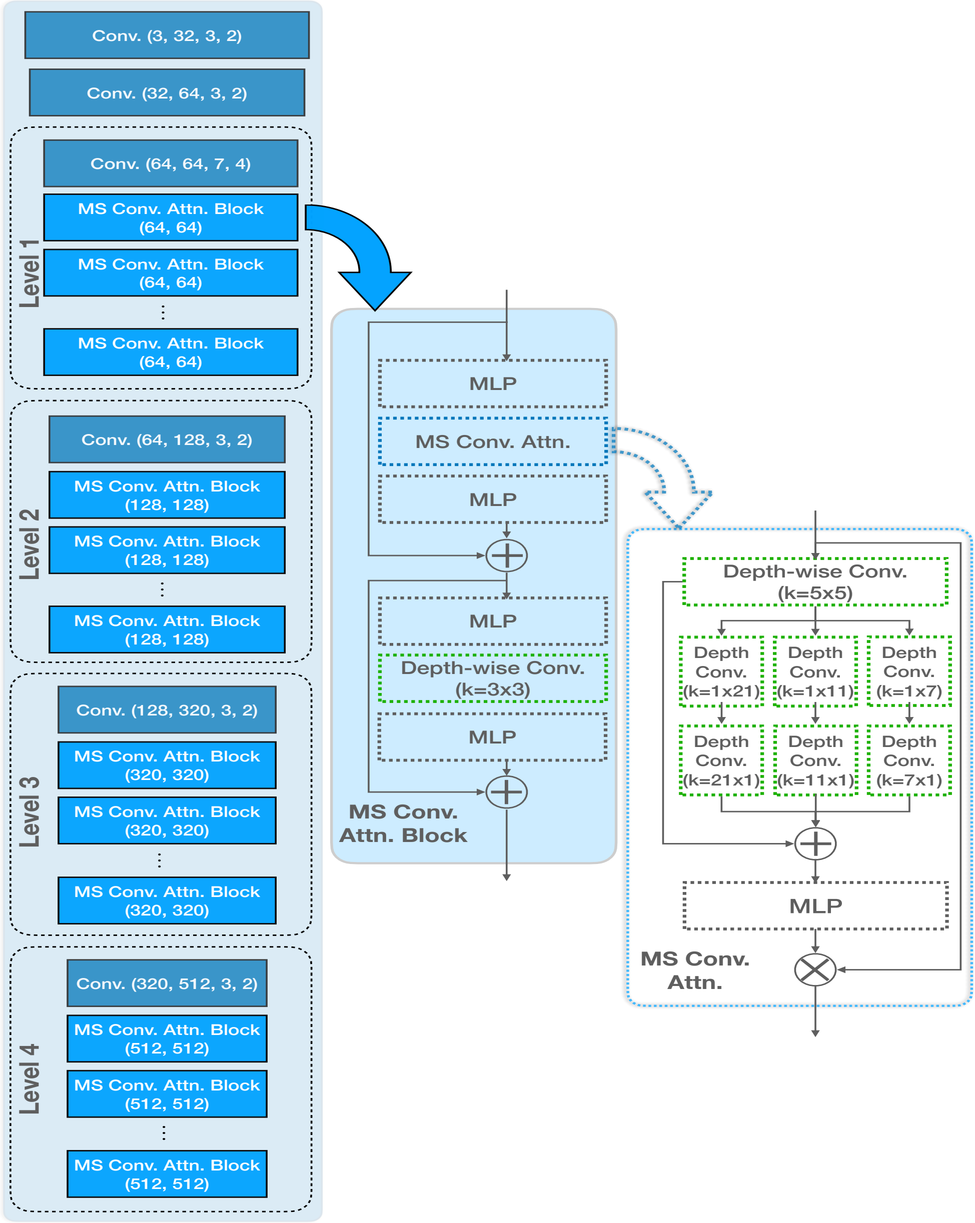

Figure 1: The framework of the four-level feature extractor in our proposed model has shown on the left. For each level, we utilize a convolutional layer, and for the four levels, we employ [3, 3, 12, 3] multi-scale convolutional attention (MS Conv. Attn.) blocks, respectively. Each multi-scale convolutional attention block (as depicted in the middle blue block) incorporates a multi-scale convolutional attention with depth-wise convolutions (as shown on the right). The numbers within each convolutional layer indicate the input channel size, output channel size, kernel size and stride. The numbers associated with each multi-scale attention block denote the input and output channel sizes of the block, indicating that the output channel size is set to match the input channel size within each block, including the layers within the multi-scale convolutional attention. The number within each depth-wise convolutional layer represents the kernel size.

attention, is the same as the number of input channels. To strike a balance between performance and complexity, we utilize [3, 3, 12, 3] multi-scale convolutional attention blocks for the four levels, respectively.

2 Details for Progressive Decoder

To effectively leverage the multi-scale features extracted from the feature extractor, we have specifically designed a lightweight progressive decoder that gradually incorporates details to achieve the final high-resolution prediction. In total, we employ three lightweight hamburger heads, as depicted in Figure 2 of the paper. Within each head (H_1~H_3), the output channel sizes are set [512, 128, 64], respectively. As for MLP_1~MLP_3, the output channel size is set to 1.

3 More Qualitative Results

In the paper, we have presented visualizations of dichotomous segmentation results obtained from our proposed model, comparing them to the ground truth and results generated by alternative approaches. In this supplementary section, we provide additional examples for further qualitative comparison, as shown in Fig. 2, Fig. 3, Fig. 4, Fig. 5. These examples include results generated by one-stage competitors PGNet [4] and IS-Net [3]. Our model consistently produces high-quality results with clearer shapes and finer details, outperforming the other methods.

Furthermore, we show examples comparing our results to those obtained by a multi-stage method InSpyReNet [2] in Fig. 6. Our method excels at recognizing sharp object details, such as the string around the windmill and the small hole between cracks. However, it may encounter challenges in cases where the object is overexposed to light, as seen with the light bulb in the fifth row example. Objects that are completely mixed with another object present a difficult task for both multi-stage and one-stage methods. For instance, our method recognizes the plates under the tomatoes as the main object (as shown in the last row of Fig. 6). Although InSpyReNet successfully segments parts of the tomatoes, it also segments the plates.

References

[1] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zheng-Ning Liu, Ming-Ming Cheng, and Shi min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. In *NeurIPS*, 2022.

[2] Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, and Daijin Kim. Revisiting image pyramid structure for high resolution salient object detection. In *ACCV*, 2022.

[3] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *ECCV*, 2022.

[4] Chenxi Xie, Changqun Xia, Mingcan Ma, Zhirui Zhao, Xiaowu Chen, and Jia Li. Pyramid grafting network for one-stage high resolution saliency detection. In *CVPR*, 2022.

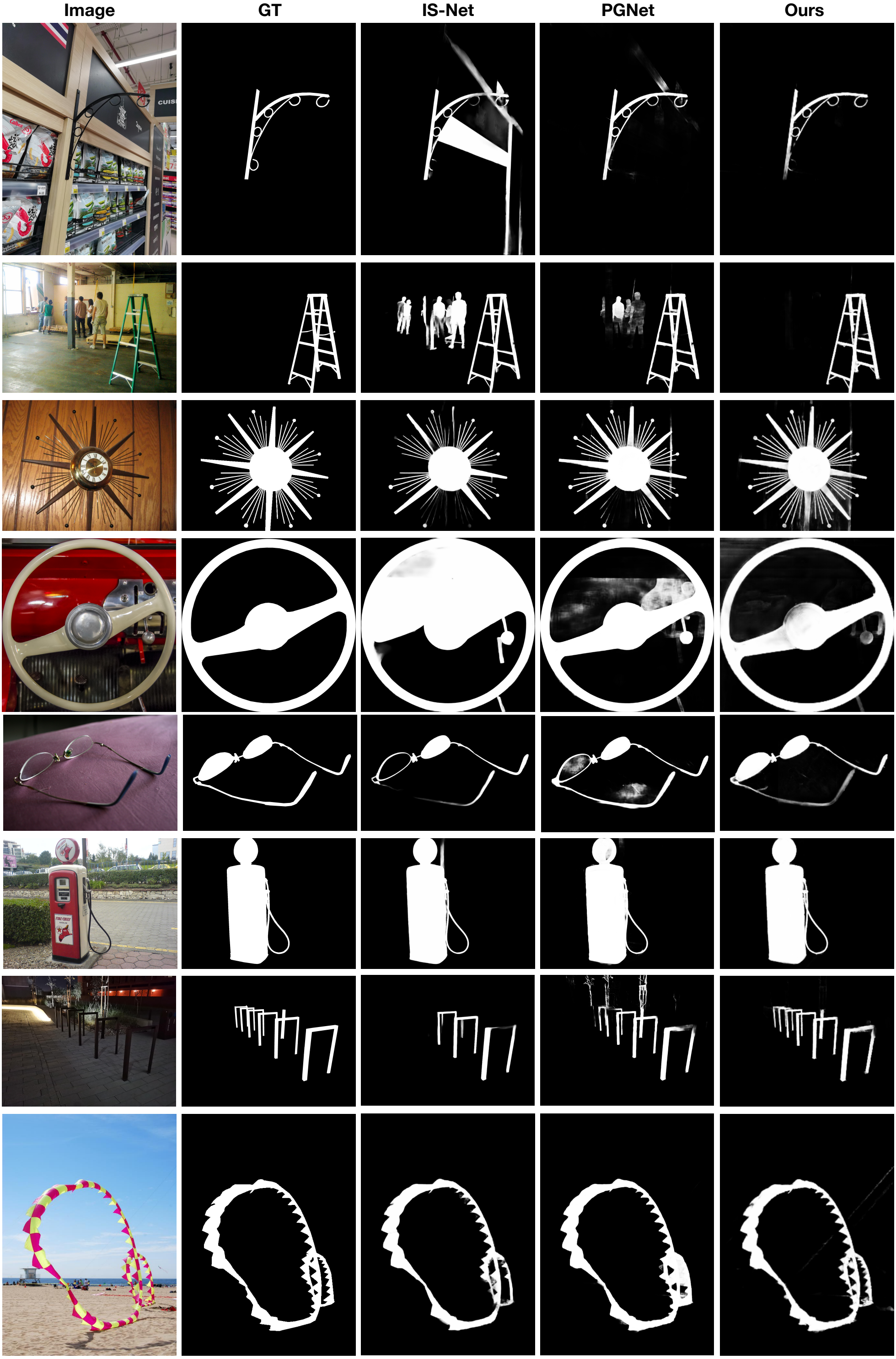

Figure 2: Qualitative comparison with the state-of-the-art one-stage methods on the DIS5K DIS-TE1 dataset. Our proposed method accurately segments main objects while preserving finer object details.

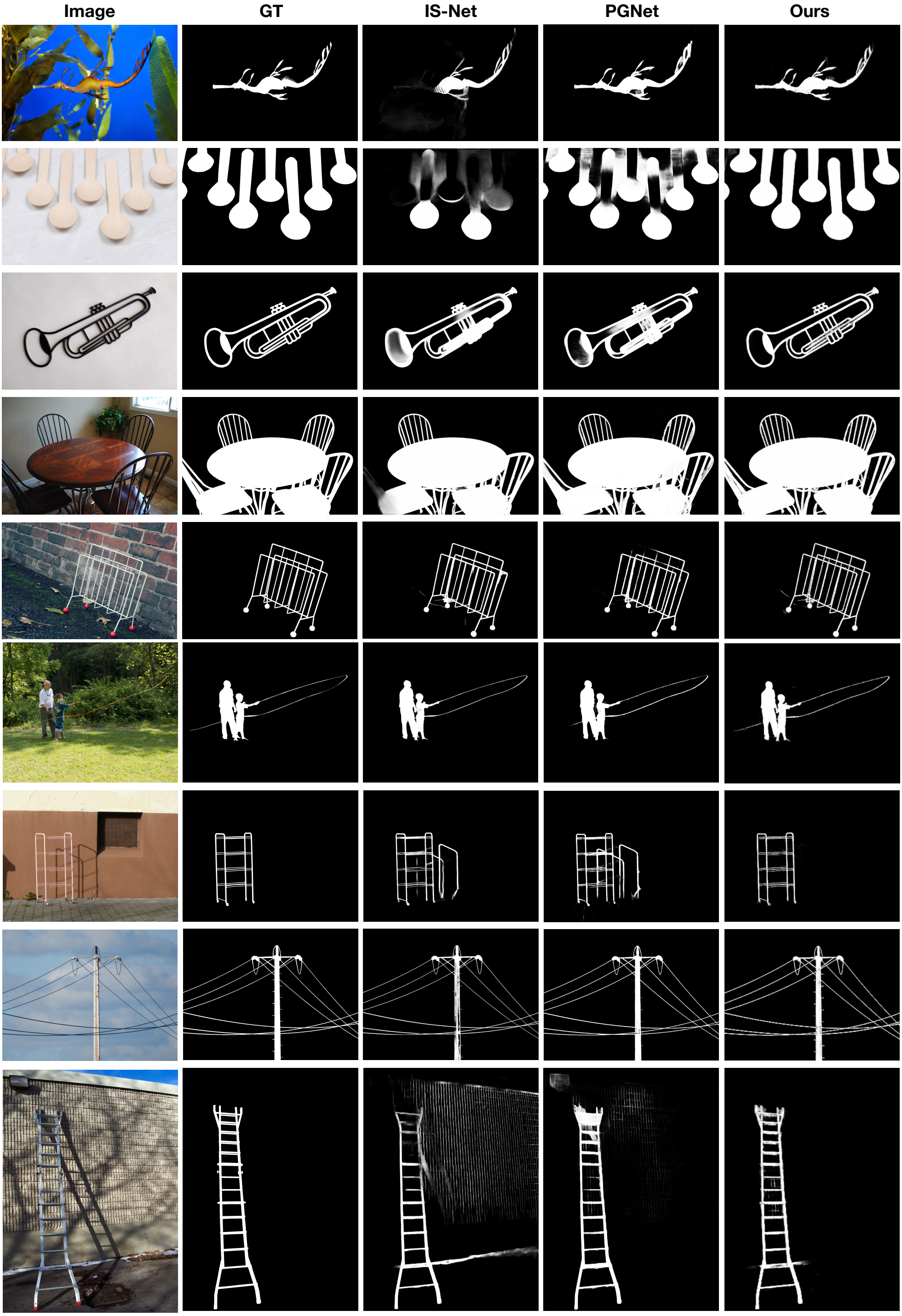

Figure 3: Qualitative comparison with the state-of-the-art one-stage methods on the DIS5K DIS-TE2 dataset. Our proposed method is capable of capturing very fine object details, such as cables and lines.

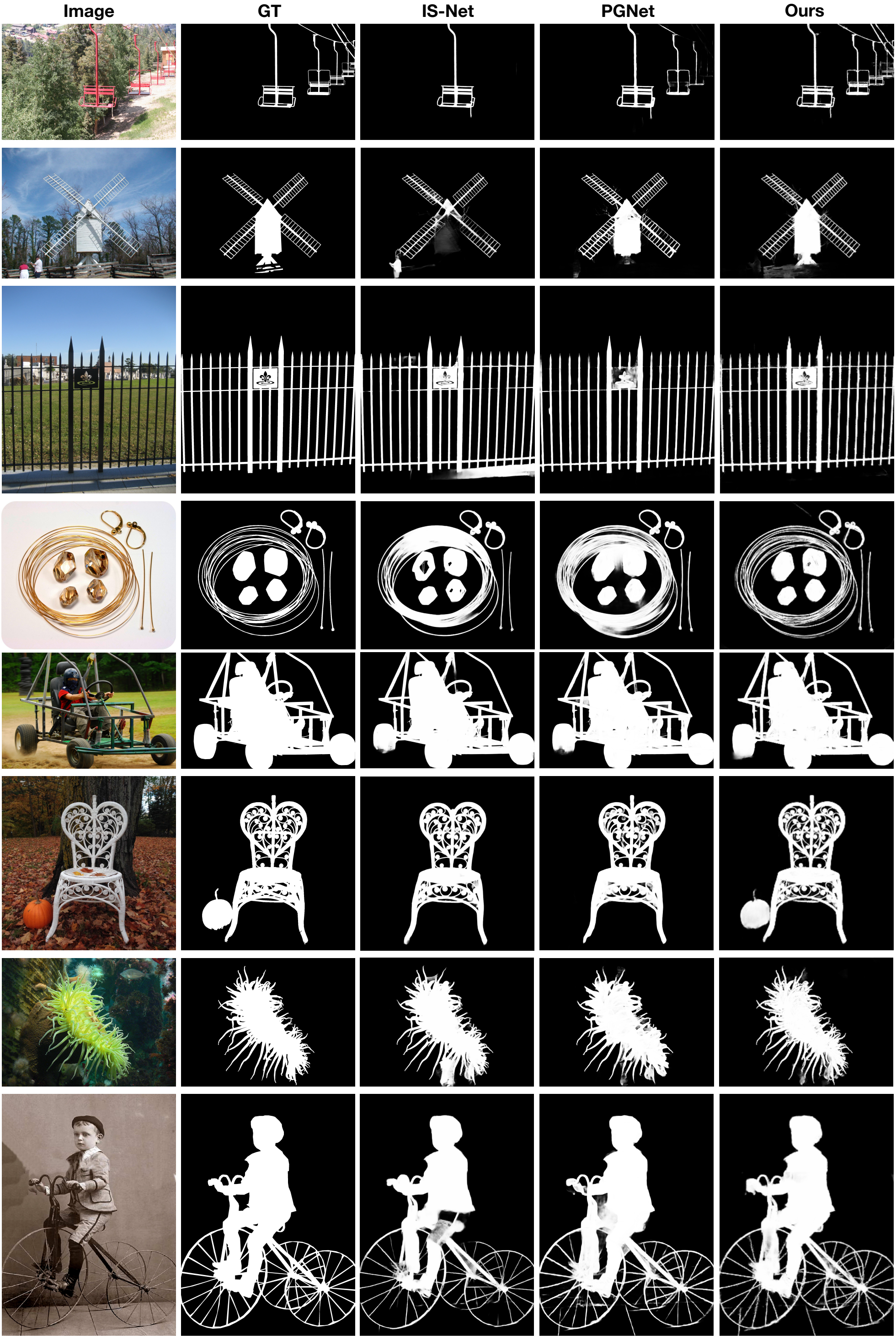

Figure 4: Qualitative comparison with the state-of-the-art one-stage methods on the DIS5K DIS-TE3 dataset. Our proposed method surpasses other methods in capturing finer object details.

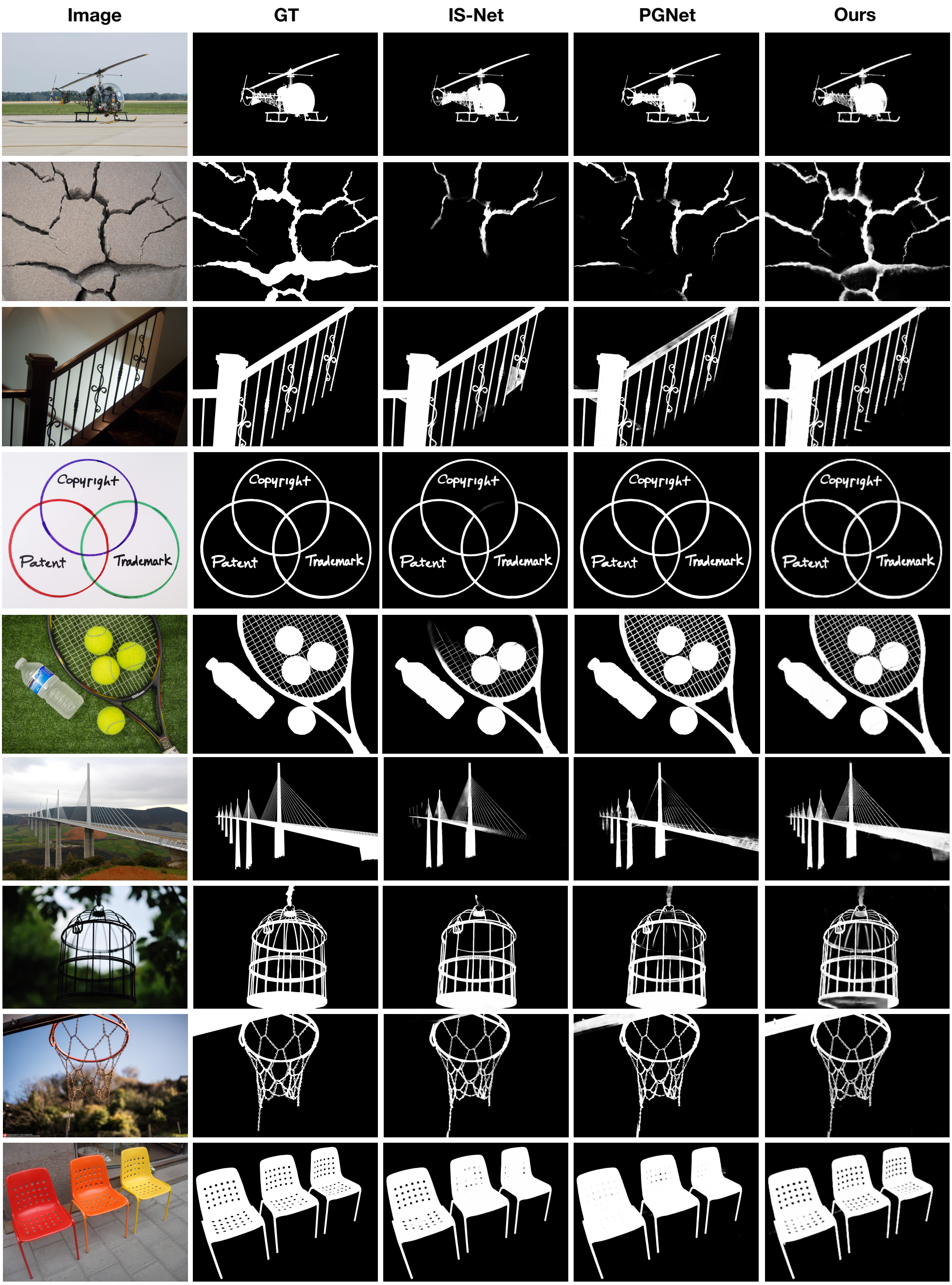

Figure 5: Qualitative comparison with the state-of-the-art one-stage methods on the DIS5K DIS-TE4 dataset. Our proposed method excels in segmenting objects with sharper edges.

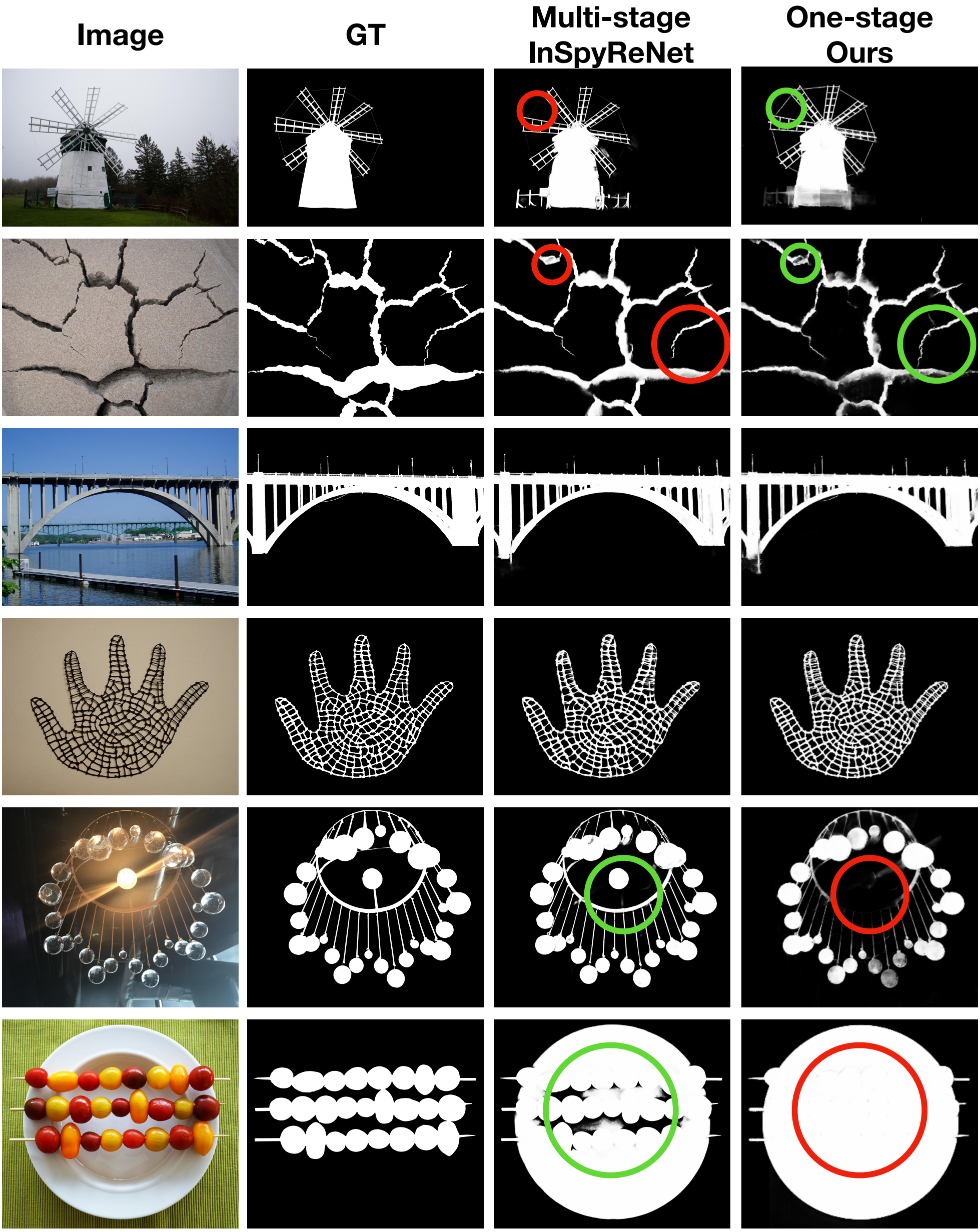

Figure 6: Qualitative comparison with the state-of-the-art multi-stage method InSpyReNet [2] on the DIS5K testing dataset. Correct predictions are marked with a green circle, while unsatisfactory areas in the predictions are marked with a red circle. The examples in the first two rows demonstrate that our method performs better in capturing sharp details, such as the string around the windmill and the small hole between cracks. In most cases, our method produces comparable results to the multi-stage method as shown in the examples in the third and fourth rows. However, our method may struggle with objects under overexposed lighting conditions, as seen in the image in the fifth row. Additionally, our method can encounter challenges when two different objects are mixed together, as demonstrated in the tomato example in the last row. While the multi-stage InSpyReNet method is able to recognize parts of the tomatoes, it also segments the plate beneath the tomatoes.