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

Abstract

Dichotomous segmentation is a challenging task that involves recognizing foreground objects in high-resolution images with varying characteristics. Existing methods often miss important details of the object or require a long processing time due to multi-stage process. In this paper, we propose a one-stage effective model that can distinguish objects in dichotomous segmentation with low computation cost. Unlike most methods that use two separate branches to first obtain coarse results from low-resolution images and then refine them with the high-resolution information, our method can directly process high-resolution inputs with simple operations. We introduce convolutional attentions into the feature extractor to effectively capture multi-scale features. These features are then used to generate high-quality results with a specifically designed progressive decoder. The experimental results demonstrate that our method achieves superior performance on the DIS5K dichotomous segmentation dataset with fewer model parameters and computational operations.

1 Introduction

Dichotomous segmentation is a recently proposed task that aims to identify foreground objects on high-resolution nature images with varying characteristics, *e.g.*, salient, common, camouflaged and meticulous. This task has high potential impacts on various computer vision tasks, such as depth estimation, 3D modeling, 3D model editing. The first work on dichotomous segmentation, IS-Net [28], has achieved impressive results on the DIS5K dataset. Although IS-Net introduces intermediate supervision on features to strengthen the model learning, the prediction still misses some sharp object details, as shown in Fig. 1. Later, InSpyReNet [20] is proposed achieving the state-of-the-art performance on both saliency detection and dichotomous segmentation with a pyramid blending schema to enable training

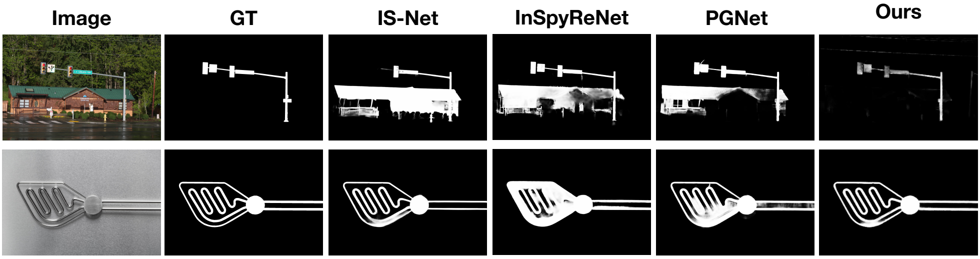

Figure 1: Comparison of different methods (*i.e.* IS-Net [28], InSyReNet [20], PGNet [32] and ours) on DIS5K dataset. The compared alternatives segment the house behind the traffic light as part of the foreground object and mix the part of the wire whisk, while our method can generate clean prediction with sharp edges.

on low-resolution images (384 × 384) while testing on high-resolution ones (1024 × 1024). However, the multi-stage process of InSpyReNet dramatically increases the processing time and the number of FLOPs for each high-resolution image. This could greatly limit its application in real-world scenarios.

Although some recent high-resolution saliency detection approaches can also be adapted to the dichotomous segmentation task, they may fail in some cases, as this task includes more diverse samples than saliency. There have been several recent high-resolution saliency detection methods that adopt a multi-stage architecture to predict coarse results from low-resolution images and then add details in high-resolution images, such as HRSOD [36] and DHQSOD [29]. However, the multi-stage process significantly slows down the prediction speed. To address this issue, PGNet [32] designs a faster one-stage model. In order to employ a high computation cost transformer, it still follows the traditional two-branch schema, where one branch is for the low-resolution input and the other is for the high-resolution one. Using a transformer might boost the performance, while it would greatly increase the size of the model and the number of operations simultaneously, making it hard to train.

In this paper, we propose an efficient model that directly works on high-resolution images without relying on low-resolution inputs. To mitigate the computation complexity, our model is designed as a one-stage encoder-decoder-structure. As the encoder generates abstract features from the image and decreases the feature map resolution, the low-resolution image features are inherently included within the encoder feature maps, eliminating the need to process extra downsampled low-resolution images like HRSOD [36], PGNet [32] or InSpyReNet [20]. Another common reason for recent methods to start from low-resolution images is the high complexity of their feature extractor, such as using transformers, which prevents directly processing high-resolution images due to the limited computation resources, *i.e.*, GPU memory. Conversely, our encoder is efficient enough to directly process high-resolution images as we introduce a convolutional attention module into the encoder that can more effectively encode contextual information than the self-attention mechanism in a transformer with less computational cost. We extract multi-scale features from the encoder at different levels and design a progressive segmentation decoder to gradually increase the resolution for the segmentation map level by level with the extracted features. To strengthen learning power, multi-scale supervision is adopted during training.

To validate our proposed model, we train and test it on the large-scale dichotomous segmentation dataset DIS5K and provide a comprehensive experimental comparison with recent methods. The experimental results show that our proposed model can predict high-quality segmentation on high-resolution images. Furthermore, our model outperforms all single-stage competitors with a relatively small number of computational operations during prediction. We also conduct ablation studies to reveal the impacts of different modules in our model.

Our main contributions are summarized as follows:

- We propose a one-stage framework with an efficient yet effective convolutional attention module that could directly work on high-resolution images for dichotomous segmentation.
- We design the progressive prediction schema into the decoder of the model which enables the gradual refinement of the segmentation map level by level. A multi-scale supervision loss is introduced to enhance the model learning.
- We verify our model on the DIS5K dataset [28], where our model achieves the best performance among all the single-stage methods with lower computation complexity.

2 Related Work

Since the dichotomous segmentation is a relatively new task, there are not many relevant works. Therefore, besides providing backgrounds of the dichotomous segmentation, we also review some recent learning-based saliency detection methods and semantic segmentation methods that could be adapted to the dichotomous segmentation task.

2.1 High-Resolution Foreground Object Segmentation

Recently proposed by Qin *et al.*, dichotomous segmentation is a new task that aims to segment highly accurate objects from natural images, covering camouflaged, salient, or meticulous objects in various backgrounds [28]. The first attempt IS-Net [28] is based on U2^2-Net [27] with intermediate supervision on the feature level, and has achieved impressive results in the large-scale dichotomous segmentation dataset DIS5K. Limited by the computation resource, many other alternative foreground object segmentation methods adopt coarse-to-fine schema that gets the global semantics on low-resolution images first and then refines the details with features from high-resolution. For example, InSPyReNet [20] trains a large model with low-resolution images and proposes a pyramid blending module that adds high-resolution details into the low-resolution prediction to get the final segmentation. The two-stage process in the InSPyReNet for high-resolution prediction is time- and memory-consuming. PGNet [32] uses a transformer to predict a saliency map from low-resolution images, and then fuses high-resolution features via a grafting module to improve the prediction. It has been demonstrated through a number of tasks that incorporating a transformer generally leads to performance improvement [6, 22, 35]. Nonetheless, this enhancement comes at the cost of increased computational demands.

Besides InSpyReNet and PGNet, HRSOD and DHQSOD are another two popular methods for high-resolution saliency detection. HRSOD [36] first extracts global features then optimizes local details by splitting images into patches and finally fuses the prediction. Tang *et al.* [29] propose the DHQSOD model to disentangle the task into classification and regression subtasks. They first design a a low-resolution saliency classification network to capture sufficient semantics at low resolution and generate the trimap. The high-resolution refinement network is then introduced with an uncertainty loss to refine the low-resolution trimap generated in the first stage to a higher resolution one. Although those multi-stage coarse-to-refine methods can produce precise predictions, they often require significant time and memory resources, which can pose challenges in certain real-world applications. Furthermore, these methods can lead to significant semantic inconsistencies between low-resolution and high-resolution networks. Therefore, our goal is to design a one-stage network that operates directly on high-resolution inputs to address the aforementioned limitations.

2.2 Semantic Segmentation

Apart from dichotomous segmentation and saliency detection, our task is also closely related to the field of semantic segmentation. The field of semantic segmentation has witnessed remarkable advancements with the introduction of fully convolutional networks [23], which employ an end-to-end per-pixel classification approach. Further improvements have been achieved by incorporating multi-scale features [5, 18], channel- and self-attention blocks [8, 14, 16, 19], and utilizing edge cues [3, 11].

Recently, there has been a surge of interest in leveraging the vision transformers in various tasks, leading to the emergence of dense prediction transformers and semantic segmentation transformers, such as SegFormer [33], TopFormer [37], MaskFormer [7]. Although vision transformers have demonstrated significant improvements in performance, they also entail a high memory overhead. Moreover, some researchers have discovered that incorporating multi-modal and multi-task data can enhance the learning capabilities of a model and yield further benefits for semantic segmentation tasks [2, 34]. However, the majority of these models are developed specifically for autonomous driving scenarios, which require at least RGB images and depth information as paired inputs. While it is possible to employ a semantic segmentation model for dichotomous segmentation, such an approach may be prone to overfitting. This is because semantic segmentation models are usually tailored for complex segmentation scenarios that involve multiple classes, whereas dichotomous segmentation mainly entails distinguishing between foreground and background.

3 Approach

To overcome the challenges associated with dichotomous segmentation, we present a one-stage learning model based on an encoder-decoder architecture. Our model can directly work on high-resolution images without requiring a separate branch for processing low-resolution inputs. Since the encoder keeps extracting abstract features from the image while simultaneously reducing the feature map resolution, low-resolution image features are inherently integrated into the encoder. We design an efficient decoder that can gradually leverage the multi-scale features from the encoder to generate the final high-resolution segmentation maps. As shown in Fig. 2, our proposed method mainly contains two components, an encoder, *i.e.*, feature extractor, and a decoder for progressive prediction. Details of the two components are further discussed below.

3.1 Feature Extractor with Multi-scale Convolutional Attention

Attention mechanism is a kind of adaptive selection process that has proved its effectiveness in a variety of computer vision tasks [10, 21, 31] by enabling the network to focus on

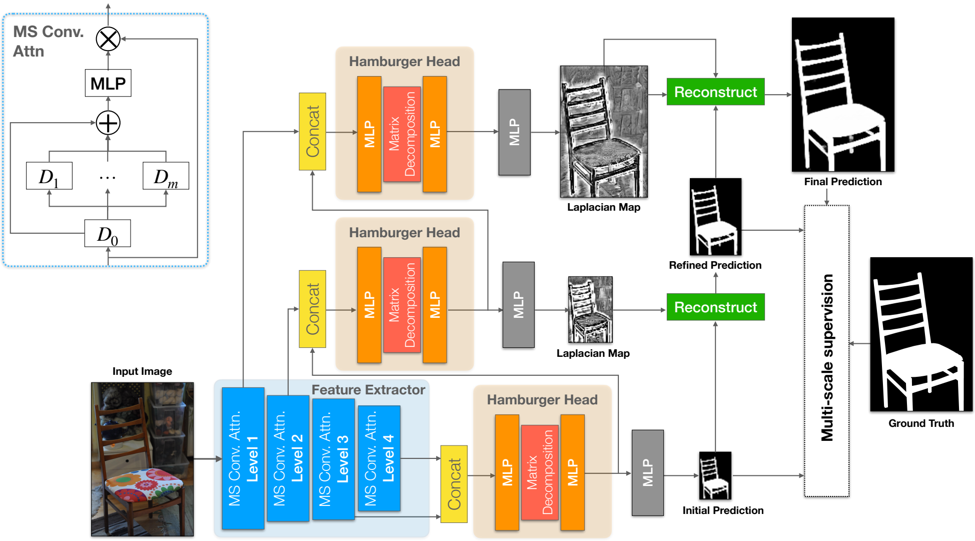

Figure 2: The framework of our proposed model, which consists of a feature extractor with multi-scale convolutioanl attentions to generate effective multi-scale features, and a progressive decoder with hamburger heads to gradually utilize the multi-scale features to achieve high-resolution results. Multi-scale supervision is introduced to enhance the model's training capabilities. An overview of the multi-scale convolution attention is depicted at the upper left, where D_0~D_m denote depth-wise convolutions. More details are presented in the supplementary material.

significant parts. Although transformers have been increasingly employed as feature extractors due to their impressive learning capabilities from self-attention mechanisms, the high computation complexity of transformers makes it unsuitable for a practical model that operates directly on high-resolution inputs. Consequently, we carefully choose a CNN-based backbone with an effective attention mechanism when designing the feature extractor.

The multi-scale convolutional attention module [16] is an effective attention-based module that includes a depth-wise convolution to aggregate local information, multi-branch depth-wise convolutions with different sizes of the receptive fields to capture multi-scale context, and a MLP layer to model relationship between different channels. The output of the MLP layer is used as attention weights directly to reweigh the input of the convolutional attention module. The covolutioanl attention has shown strong impact on the semantic segmentation task, outperforming transformers with less computational cost [16]. We construct a four-level feature extractor with each level containing multiple multi-scale convolutional attention modules (as depicted at the upper left of Fig. 2) and a convolutional layer to get features with decreasing spatial resolution. More details of the multi-scale convolutional attention are presented in the supplementary material.

3.2 Progressive Decoder

To effectively leverage the multi-scale features obtained from the feature extractor, we design a progressive decoder with heads for each scale, which generates an initial prediction using the lowest-level features and then gradually add details from the higher-level features while also increasing the resolution to produce the final prediction.

As shown in Fig. 2, the initial prediction is obtained upon the features concatenated by those from level 3 and 4. Specifically, we adopt a light yet powerful hamburger head [15] that consists of a matrix decomposition between two MLP layers, as its performance surpasses various decoder heads with only O(n) complexity. Let F_i denote the output multi-scale features from the i-th level of the feature extractor, H_1 be the first light hamburger head, and MLP_1 be the MLP layer after the first hamburger head, the initial dichotomous segmentation result ($R_{initial}$) can be computed as

$$F_{H_1} = H_1(Concat(F_3, F_4)), \quad R_{initial} = MLP_1(F_{H_1}). \tag{1}$$

After that, the initial prediction result will be refined by a reconstruction operation together with a Laplacian map generated by another hamburger head (H_2) and a MLP layer (MLP_2) from the concatenated features of F_2 and F_{H_1}. To address the resolution difference, F_{H_1} will be upsampled with bilinear interpolation before concatenation. The Laplacian map ($LP_{refined}$) here should contain high-frequency details.

$$F_{H_2} = H_2(Concat(F_2, F_{H_1})), \quad LP_{refined} = MLP_2(F_{H_2}). \tag{2}$$

Subsequently, we employ the Laplacian map to reconstruct a refined high-resolution prediction using the initial prediction $R_{initial}$ which needs to be upgraded to match the resolution of the Laplacian map $LP_{refined}$ by first filling zeros to the empty space of a larger-size map and then calculating the convolution values with Gaussian weights [4]. The refined result can be described as

$$R_{refined} = Upgrade(R_{initial}) + LP_{refined}. \tag{3}$$

The final prediction can be obtained following a similar process of the refined prediction generation, *i.e.*, Eq.(2) and Eq.(3). First, we collect a Laplacian map LP_{final} that comes from the hamburger head (H_3) giving the concatenation of F_1 and F_{H_2}. Then the refined prediction ($R_{refined}$) is upgraded, and the values from the Laplacian map and the upgraded refined prediction are summed to yield the final results as follows:

$$\begin{aligned} F_{H_3} = H_3(Concat(F_1, F_{H_2})), \quad LP_{final} = MLP_3(F_{H_3}), \\ R_{final} = Upgrade(R_{refined})) + LP_{final}. \end{aligned} \tag{4}$$

3.3 Training Loss

To enhance the learning capabilities of the model, we enforce supervision on the multi-scale predictions obtained from different levels. Since the dichotomous segmentation involves only two categories in the segmentation map, we utilize the classic binary cross entropy (BCE) loss in our model. The loss function can be formulated as

$$L = \text{argmin } \lambda_1 BCE(R_{initial}, G) + \lambda_2 BCE(R_{refined}, G) + \lambda_3 BCE(R_{final}, G), \tag{5}$$

where G is the ground truth dichotomous segmentation map, and λ_1, λ_2 and λ_3 are the weighting hyperparameters.

4 Experiments

In this section, we present an overview of the dataset DIS5K, model implementation details, the experimental outcomes with comparison to the existing methods, and the ablation studies to show the impact of each module.

4.1 Dataset and Evaluation Metrics

Since the dichotomous segmentation is a relatively new task, to the best of our knowledge, DIS5K is currently the only dataset available for this task. Collected by Qin *et al.* [28], DIS5K includes 5,470 Flickr images covering 225 categories in 22 groups. The images primarily contain individual foreground objects with intricate and highly precise structures and details, irrespective of their saliency, commonality, camouflage, meticulousness, and other attributes. The 5,470 images in DIS5K are divided into three subsets: 3,000 images in DIS-TR for training, 470 images in DID-VD for validation and 2,000 images in DIS-TE for testing. The 2,000 images of the DIS-TE are further divided into four subsets (DIS-TE1 ~DIS-TE4), each containing 500 images.

In accordance with [28], a total of six metrics have been utilized to comprehensively evaluate the performance from different perspectives. The metrics include maximal F-measure (F^{mx}) [1], weighted F-measure (F^{w}) [24], mean absolute error (M) [26], structural measure (S) [12], mean enhanced alignment measure (E) [13] and human corrections efforts (HCE) [28]. In the case of F^{mx}, F^{w}, S, and E, a higher score indicates better performance. Conversely, for M and HCE, a lower score suggests better performance.

4.2 Implementation Details

We train our model with the training subset (DIS-TR) of DIS5K, validate on the DIS-VD subset and report the performance on the four testing subsets (DIS-TE1 ~DIS-TE4). In order to balance the performance and the complexity, we use $[3,3,12,3]$ multi-scale convolutional attention modules for the four levels respectively. The kernel size of the depth-wise convolution layers within each attention modules are set to 1x7, 7x1, 1x11, 11x1, 1x21, 21x1. The feature extractor is pretrained on ImageNet-1K dataset [9] before training on the DIS5K. We also adopt some common data augmentation including random horizontal flipping, random scaling (from 0.5 to 2) and random cropping for training. The batch size is set to 2. AdamW [17] is used as the optimizer. We set the initial learning rate as 0.00006, and use the poly learning rate decay for scheduling [38]. We implement our model by using PyTorch and train it in 152K iterations with input size 1024×1024. λ_1, λ_2 and λ_3 in Eq. (5) are set to 1.

4.3 Quantitative and Qualitative Results

We compare our proposed model with 7 recent deep-learning-based models designed for different segmentation tasks, including high resolution saliency detection models (HRSOD [36], PGNet [32] and InSpyReNet [20]), camouflaged object segmentation model (PFNet [25]), semantic segmentation models (HRNet [30] and SegNeXt [16]) and dichotomous segmentation model (IS-Net [28]). The metric scores were obtained from models trained with the same DIS-TR training set as our approach. The inference time was calculated using the same machine with a Tesla V100 GPU for a fair comparison.

Table 1 presents the quantitative results, indicating that our proposed approach significantly outperforms all the single-stage methods including the dichotomous segmentation method IS-Net. Furthermore, our method can handle high-resolution inputs with fewer model parameters and less computational operations. It is evident from the table that multi-stage models generally require more time for prediction. Among all the multi-stage methods, InSpyReNet produces the best scores but also has the highest inference time due to its multi-stage design. Our model performs second-best compared to the multi-stage methods with

Table 1: Comparison of dichotomous segmentation on four DIS5K [28] testing subsets with alternative approaches. Overall performance is computed by taking the mean value of the scores from the four subsets. The best performance has been **bolded** and the second best result is marked in **blue**. Higher F^{mx}, F^{w}, S, E scores and lower M, HCE values indicate the better performance. Our method outperforms all the the single-stage methods with fewer model parameters and computational operations (FLOPs).

Dataset	Metric	Multi-stage		Single-stage					
		HRSOD [36]	InSpyReNet [20]	PFNet [25]	HRNet [30]	SegNeXt [16]	IS-Net [28]	PGNet [37]	Ours
	Params(M)	32.4	90.7	46.5	63.6	**27.6**	44.0	72.7	**28.0**
	Time (ms)	425.7	733.0	**70.5**	172.6	275.7	**80.5**	127.8	287.0
	FLOPs (G)	315.7	461.2	**59.9**	373.8	137.6	159.8	160.3	**129.2**
	Input Size	1024^2	1024^2	416^2	1024^2	1024^2	1024^2	1024^2	1024^2
DIS-TE1	$F^{mx}\uparrow$	0.726	**0.854**	0.646	0.668	0.771	0.740	0.821	**0.822**
	$F^{w}\uparrow$	0.658	**0.792**	0.552	0.579	0.681	0.662	0.728	**0.745**
	$M\downarrow$	0.079	**0.044**	0.094	0.088	0.076	0.074	0.070	**0.069**
	$S\uparrow$	0.766	**0.873**	0.722	0.742	0.789	0.787	0.834	**0.845**
	$E\uparrow$	0.803	**0.893**	0.786	0.797	0.820	0.820	0.846	**0.859**
	$HCE\downarrow$	198	**110**	253	262	177	149	173	**147**
DIS-TE2	$F^{mx}\uparrow$	0.781	**0.895**	0.720	0.747	0.826	0.799	0.841	**0.857**
	$F^{w}\uparrow$	0.714	**0.846**	0.633	0.664	0.741	0.728	0.782	**0.802**
	$M\downarrow$	0.074	**0.038**	0.096	0.087	0.068	0.070	**0.066**	**0.066**
	$S\uparrow$	0.795	**0.905**	0.761	0.784	0.828	0.823	0.842	**0.867**
	$E\uparrow$	0.832	**0.925**	0.829	0.840	0.879	0.858	0.888	**0.903**
	$HCE\downarrow$	467	**255**	567	555	427	**340**	405	346
DIS-TE3	$F^{mx}\uparrow$	0.806	**0.912**	0.751	0.784	0.843	0.830	0.877	**0.882**
	$F^{w}\uparrow$	0.732	**0.868**	0.664	0.700	0.765	0.758	0.803	**0.831**
	$M\downarrow$	0.069	**0.038**	0.092	0.080	0.062	0.064	0.059	**0.058**
	$S\uparrow$	0.819	**0.915**	0.777	0.805	0.841	0.836	0.857	**0.873**
	$E\uparrow$	0.863	**0.942**	0.854	0.869	0.899	0.883	0.906	**0.912**
	$HCE\downarrow$	1007	**523**	1082	1049	871	687	838	**680**
DIS-TE4	$F^{mx}\uparrow$	0.789	**0.902**	0.731	0.772	0.834	0.827	0.859	**0.863**
	$F^{w}\uparrow$	0.726	**0.847**	0.647	0.687	0.755	0.753	0.798	**0.819**
	$M\downarrow$	0.072	**0.046**	0.107	0.092	0.069	0.072	0.067	**0.066**
	$S\uparrow$	0.804	**0.902**	0.763	0.792	0.823	0.830	0.844	**0.870**
	$E\uparrow$	0.848	**0.927**	0.838	0.854	0.883	0.870	0.895	**0.908**
	$HCE\downarrow$	3720	**2336**	3803	3864	3679	2888	3449	**2768**
Overall	$F^{mx}\uparrow$	0.776	**0.890**	0.712	0.743	0.818	0.799	0.849	**0.856**
	$F^{w}\uparrow$	0.708	**0.838**	0.624	0.658	0.735	0.726	0.778	**0.799**
	$M\downarrow$	0.074	**0.042**	0.097	0.087	0.069	0.070	0.065	**0.064**
	$S\uparrow$	0.796	**0.898**	0.756	0.781	0.820	0.819	0.844	**0.864**
	$E\uparrow$	0.837	**0.922**	0.827	0.840	0.870	0.858	0.884	**0.896**
	$HCE\downarrow$	1348	**806**	1427	1432	1289	1016	1216	**986**

Table 2: The ablation studies on DIS-TE1 to verify the effectiveness of each module. We can observe the progressive schema has a greater impact on the model's performance than the multi-scale supervision, while they together yield the best performance.

Decoder		Supervision							
Single Head	Progressive	Single-scale	Multi-Scale	$F^{mx}\uparrow$	$F^{w}\uparrow$	$M\downarrow$	$S\uparrow$	$E\uparrow$	$HCE\downarrow$
✓		✓		0.758	0.676	0.079	0.772	0.813	211
	✓	✓		0.806	0.729	0.073	0.817	0.834	169
	✓		✓	**0.822**	**0.745**	**0.069**	**0.845**	**0.859**	**147**

nearly one-third of the inference time of InSpyReNet, offering a better balance between performance and speed.

Besides, we visualize the dichotomous segmentation predictions from all the competitors and list some examples in Fig. 1 and Fig. 3. Our model effectively capture finer object details, such as chair legs and octopus legs. However, it may struggle with objects have reflection and are closely aligned with other objects, such as the flagpole in the third row of Fig. 3. Though InSpyReNet predicts the contour of the flagpole, it also segments the pipe object in the building behind it. More results can be found in the supplementary material.

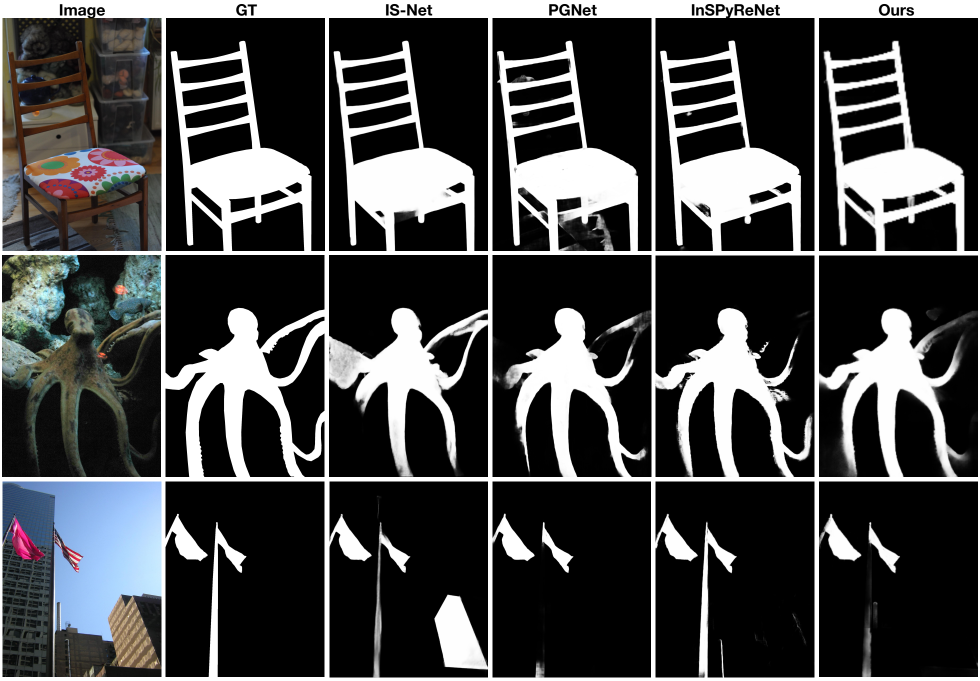

Figure 3: Qualitative comparison with the state-of-the-art methods on DIS5K dataset. Our proposed method is capable of capturing finer object details, such as chair legs and octopus legs. However, it may struggle in cases where an object has a reflection and is closely aligned to another object, such as the flagpole in the third row. Though InSpyReNet[20] predicts the contour of the flagpole, it also segments the pipe object in the building behind it.

4.4 Ablation Studies

We conducted ablation studies to investigate the impact of the progressive decoder and the multi-scale supervision. Firstly, instead of using the progressive decoder, we use one hamburger head with the concatenation of the features from level 3 and 4 in the feature extractor to obtain a prediction (similar to the initial prediction shown in the Fig. 2). Secondly, we add the progressive decoder but train the model with single-scale supervision only on the final prediction. Table 2 shows the metric scores from models with different modules, revealing that the progressive schema has a greater impact on the model's performance than the multi-scale supervision, while they together yield the best performance.

5 Conclusion

In this paper, we focus on the new and valuable dichotomous segmentation task and present an efficient one-stage model that works directly on high-resolution images without any downsampled auxiliary inputs. We introduce a multi-scale convolutional attention module to effectively capture multi-scale features with low computation cost, which are used to gradually refine predictions with a progressive decoder. Multi-scale supervision is adopted to strengthen the model learning. Experiments on DIS5K dataset demonstrate that our model outperforms all single-stage competitors with less computation complexity. Ablation studies suggest the effectiveness of both the multi-scale supervision and progressive decoder.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022.

[3] Shubhankar Borse, Ying Wang, Yizhe Zhang, and Fatih Porikli. Inverseform: A loss function for structured boundary-aware segmentation. In *CVPR*, 2021.

[4] Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in Computer Vision*. 1987.

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2017.

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. In *CVPR*, 2023.

[7] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*, 2021.

[8] Sungha Choi, Joanne T Kim, and Jaegul Choo. Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In *CVPR*, 2020.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[10] Shengheng Deng, Zhihao Liang, Lin Sun, and Kui Jia. VISTA: boosting 3d object detection via dual cross-view spatial attention. In *CVPR*, 2022.

[11] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *ICCV*, 2019.

[12] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017.

[13] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, 2018.

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

[15] Zhengyang Geng, Meng-Hao Guo, Hongxu Chen, Xia Li, Ke Wei, and Zhouchen Lin. Is attention better than matrix decomposition? In *ICLR*, 2021.

[16] 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.

[17] Junjun He, Zhongying Deng, Lei Zhou, Yali Wang, and Yu Qiao. Adaptive pyramid context network for semantic segmentation. In *CVPR*, 2019.

[18] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *CVPR*, 2020.

[19] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019.

[20] 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.

[21] Dmytro Kotovenko, Pingchuan Ma, Timo Milbich, and Björn Ommer. Cross-image-attention for conditional embeddings in deep metric learning. In *CVPR*, 2023.

[22] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *ECCV*, 2022.

[23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[24] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, 2014.

[25] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, 2021.

[26] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012.

[27] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern recognition*, 2020.

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

[29] Lv Tang, Bo Li, Yijie Zhong, Shouhong Ding, and Mofei Song. Disentangled high quality salient object detection. In *ICCV*, 2021.

[30] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2020.

[31] Wangmeng Xiang, Chao Li, Biao Wang, Xihan Wei, Xian-Sheng Hua, and Lei Zhang. Spatiotemporal self-attention modeling with temporal patch shift for action recognition. In *ECCV*, 2022.

[32] 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.

[33] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 2021.

[34] Hanrong Ye and Dan Xu. Inverted pyramid multi-task transformer for dense scene understanding. In *ECCV*, 2022.

[35] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022.

[36] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *ICCV*, 2019.

[37] Wenqiang Zhang, Zilong Huang, Guozhong Luo, Tao Chen, Xinggang Wang, Wenyu Liu, Gang Yu, and Chunhua Shen. Topformer: Token pyramid transformer for mobile semantic segmentation. In *CVPR*, 2022.

[38] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.