SDETR: ATTENTION-GUIDED SALIENT OBJECT DETECTION WITH TRANSFORMER

Guanze Liu[1], *Bo Xu*[1], *Han Huang*[1], *Cheng Lu*[2], *Yandong Guo*[1,*]

[1]OPPO Research Institute, [2]Xmotors

yandong.guo@live.com

ABSTRACT

Most existing CNN-based salient object detection methods can identify fine-grained segmentation details like hair and animal fur, but often mispredict the salient object due to lack of global contextual information caused by locality convolution layers. The limited training data of the current SOD task adds additional difficulty to capture the saliency information. In this paper, we propose a two-stage predict-refine SDETR model to leverage both benefits of transformer and CNN layers that can produce results with accurate saliency prediction and fine-grained local details. We also propose a novel pre-train dataset annotation COCO SOD to erase the overfitting problem caused by insufficient training data. Comprehensive experiments on five benchmark datasets demonstrate that the SDETR outperforms state-of-the-art approaches on four evaluation metrics, and our COCO SOD can largely improve the model performance on DUTS, ECSSD, DUT, PASCAL-S datasets.

Index Terms— salient object detection, transformer on dense prediction, COCO SOD dataset

1. INTRODUCTION

Salient object detection is a basic computer vision task that aims to detect the objects in an image that attracts human attention [1]. Recent studies utilize Convolutional Neural Networks (CNN) for salient object detection and achieve remarkable results. CNN-based models adopt an encoder-decoder architecture to fuse multi-scale features on different semantic levels which is essential to grasp the saliency information.

However, there still remain two challenges to achieve accurate SOD prediction. First, CNN models can generate fine-grained detailed local saliency prediction like animal fur, they require deeper layers to achieve larger receptive fields, which leads to an inevitable structure information loss, like object boundaries in Fig. 1. Although, transformer-based backbones[2] can capture the non-local interaction across image patches unlike CNN layers, but the high computation cost of self-attention mechanism denies transformer from pixel-wise dense prediction.

Second, the existing SOD dataset like DUTS [3] are data-insufficient due to the high labeling cost of the pixel-level saliency map. SOD models that trained from scratch often suffer from severe overfitting problem, and fail to perform well on complex scenes. Previous methods often utilize an ImageNet pre-trained classification backbone to erase such problems.

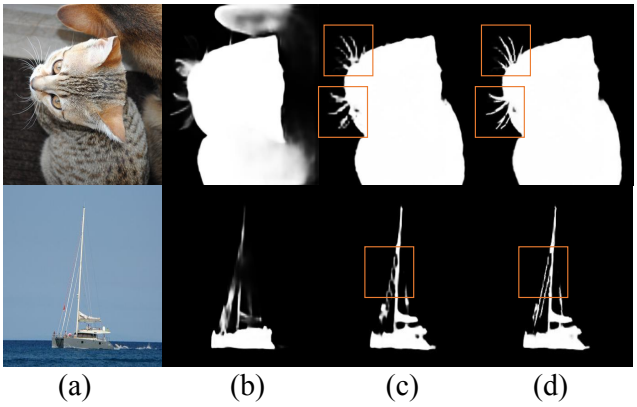

Fig. 1. Comparison results of different methods: (a) RGB image, (b) the results by F^3Net, (d) the results of our SDETR first stage (d) the results of SDETR second stage. Clearly, our SDETR achieve state of the art performance on both saliency-prediction and pixel-wise details.

To address the above challenges, we aim to propose a large pre-train dataset that's suitable for model to learn rich semantic information for saliency prediction. As shown in Figure2, we annotate the foreground object classes on the COCO panoptic segmentation dataset and produce a novel annotation called COCO SOD with more than 30000 images.

We also propose a predict-refine SDETR model to jointly combine the advantages of CNN backbone and transformer. We first predict the coarse saliency map and error map at lower resolution. A patch-level refinement network takes the low-resolution saliency map and RGB image to refine patches based on predicted error locations. As shown in Figure 1, our predict-refine SDETR achieve state-of-the-art performance in both saliency-prediction and pixel-wise details.

In summary, the contributions of this paper are threshold: (1) We propose an efficient hierarchical Transformer backbone for salient object detection. (2) We use a patch-based

refinement network to produce detailed salient object detection results. (3) We propose a novel pretrain dataset annotation COCO SOD for SOD task.

2. RELATED WORKS

2.1. Salient Object Detection

Early salient object detection methods are mainly based on hand-crafted features like color contrast, texture and certain prior to extracting saliency information, which focuses on low-level information. Recently, CNNs[4] have been widely used to extract multi-level information to produce the saliency maps.

Previous works focus on designing an effective decoder to fuse multi-level features to obtain rich semantic information for accurate salient object detection. We *et al* [5] propose F^3Net that includes a cross-feature module and a cascaded feedback decoder to fuse multi-level image features. Qin *et al* [1] introduce a boundary-aware refinement network, which is capable to produce saliency maps with sharp boundaries. Qin *et al* [4] further propose a U^2-Net architecture to capture more contextual information from different scales.

Large receptive fields are required to obtain reasonable contextual information for accurate salient object detection. Experiment results have shown that the errors on salient object detection are mainly caused by mispredicting the salient objects in the image.

2.2. Transformers on Dense Prediction.

Due to the recent success of transformer on vision tasks[6, 7], there has been a surge of interest to introduce Transformer to dense prediction tasks like semantic segmentation. Zheng *et al* [6] propose SETR, which adopts ViT as backbone and test reconstruction results with different design CNN decoders. Wang *et al* [7] proposed a pyramid vision transformer(PVT), show promising results on semantic segmentation. A Significant amount of efforts have been made to improve the performance of transformer on dense prediction tasks, such as Swin[8], Twins[9], yet pixel-wise dense prediction still remains a difficult task for transformers.

3. COCO SOD DATASET

Gathering rich, high-quality training sets has been a critical catalyst for salient object detection. Due to the high cost of pixel-wise saliency map annotation, the commonly-used SOD training-set DUTS-TR[3] is data-insufficient and contains only 10553 images in total.

In this section, we introduce our COCO SOD dataset, along with our annotation scheme. An illustration of our COCO SOD is shown in Fig.2. We generate our saliency map annotation based on panoptic segmentation labels of MS COCO[10].

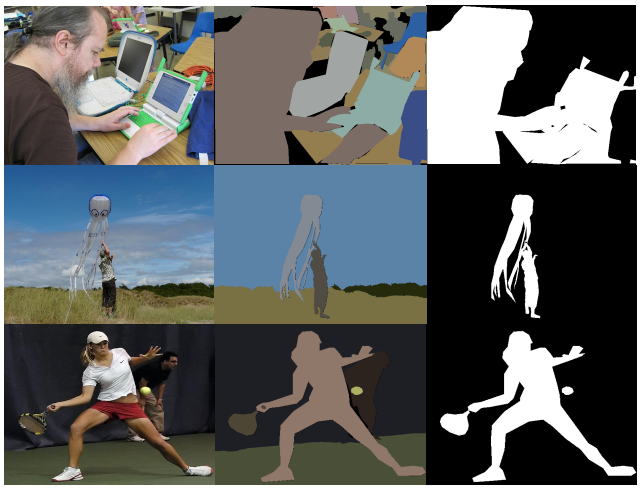

Fig. 2. Illustration of our COCO SOD dataset annotation.

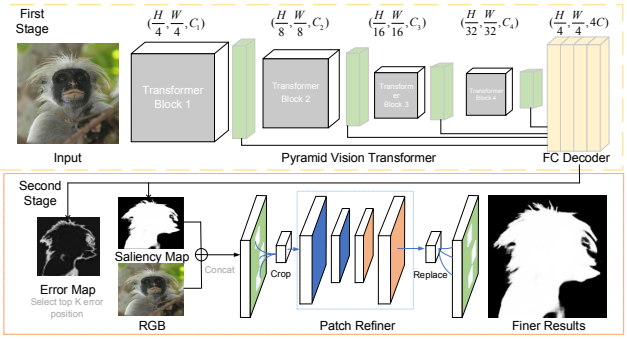

Fig. 3. Overview of the proposed SDETR network.

We select the images with a few foreground objects, and human annotators select the panoptic segmentation labels for the corresponding salient objects. The panoptic segmentation labels are converted into saliency map annotations as a two-class foreground and background segmentation label. The proposed COCO SOD dataset has more than 35000 training images, which is much larger than the existing SOD dataset.

4. METHODOLOGY

This section introduces our proposed predict-refine model, SDETR.

4.1. Transformer Encoder

4.1.1. Pyramid Vision Transformer

The purpose of a pyramid vision transformer backbone is to obtain multi-level feature maps from the input image. Both high-resolution features and low-resolution fine-grained features are required for meaningful salient object segmentation.

Given an input image of size $H \times W \times 3$, we perform overlap patch merging to progressively reduce feature map resolution from $\frac{H}{4} \times \frac{W}{4}$ to $\frac{H}{32} \times \frac{W}{32}$, each patch merging generated feature map F_i is followed by an efficient transformer block for feature learning.

4.1.2. Efficient Self-Attention.

The frequently used self-attention mechanism first computes the Q, K, V of the visual patch sequences $N = H \times W$ with the same dimensions $N \times C$. Q, K, V indicates the query, key, value of the self-attention mechanism. Self attention is denoted as:

$$Attention(Q, K, V) = Softmax(\frac{QK^T}{\sqrt{d_{head}}})V \quad (1)$$

To reduce the high computational complexity of vanilla self attention, we use sequence reduction process introduced in [7]. A reduction ratio R is introduced to reduce the length of the sequence by a factor of R times.

$$\begin{aligned} \hat{K} &= Reshape(\frac{N}{R}, C \cdot R)(K) \\ K &= Linear(C \cdot R, C)(\hat{K}) \end{aligned} \quad (2)$$

where K is the input patch sequence. Efficient self-attention first reshape feature sequence K to the one with shape of $\frac{N}{R} \times (C \cdot R)$. We reduce the complexity of the self-attention mechanism from $C \cdot R$ dimensions to C by a MLP network. We set the reduce ratio R to [64, 16, 4, 1] from stage-1 to stage-4.

We use a lightweight full MLP decoder to fuse multi-resolution features. Multi-level features F_i are first aligned to the same feature dimension using a MLP layer. Then, features with the different resolutions are resized to 1/4 of the original image resolution and concatenated. The fused features are fed to two independent MLP heads to separately predict coarse saliency map $\hat{S}_c$ and error map E_c. The Error map and coarse saliency maps are further processed by the Refine Module.

4.2. Refine Module

Due to high computational cost, our transformer backbone can only recover the saliency map to 1/4 of the original image size. A bilinear interpolation is used to preliminarily recover the full image-size saliency and error maps. However, the saliency maps predicted by the first stage has imprecise prediction, especially around instance boundaries. We use a refinement network on patches selected based on the value on the error prediction map E_c.

We select the top K error locations on the predicted error map E_c, and crop patches of size 16×16 on the corresponding locations of the coarse saliency map $\hat{S}_c$ and RGB image. The cropped saliency patches are concatenated with the corresponding image patches and then refined by refine module to generate a finer saliency map $\hat{S}_f$ to replace the origin coarse saliency map.

5. EXPERIMENT

5.1. Datasets

Training Datasets: We first pretrain our SDETR on COCO SOD, which contains more than 35000 images. We fine-tune our network on **DUTS-TR** dataset, which is a part of the DUTS[3] dataset. **DUTS-TR** contains 10553 images in total, which is the current largest salient object detection dataset with accurate pixel-wise annotations.
Evaluation Datasets: We evaluate our SDETR on five frequently used dataset benchmark including: DUT-OMRON[12], DUTS-TE[3], HKU-IS[13], ECSSD[11], PASCAL-S[14].
Evaluation Metrics: To evaluate the performance of our SDETR model, we use four evaluation metrics to measure the performance. We report Mean Absolute Error M, Mean F-measure F_β, Mean E-measure E_ξ and S-measure S_α.

5.2. Implementation Details

We train our model using state-of-the-art dense transformer backbone[20], initialized with weights trained on semantic segmentation task. We resize all the training images to size 352×352. We pretrain our SDETR on COCO SOD for 30 epochs, and train our SDETR for 60 epochs with an initial learning rate of 5×10^{-5}.

5.3. Ablation Study

We conduct a list of ablation studies on the DUTS-TE dataset[3]. We first evaluate the effectiveness of our transformer backbone in Table 2. We test four SDETR transformer blocks with four different settings, where the transformer layers of each transformer block are shown in Table 2. The results in Table 2 show that excessive transformer layers may regress the performance after its depth reaches saturation.

We then evaluate the effectiveness of our MLP decoder, by comparing the results with two commonly used decoder PSP[21] and ASPP[22] in Table3. The results on F_β, E_ξ, M demonstrate the simplicity and effectiveness of an all-MLP based decoder architecture.

We also report the results with and w/o Refine Module in Table 4. As shown in Figure 1, the results with Refine Module estimate high quality boundaries.

5.4. Comparison with State-of-the-arts

Quantitative Comparison. To demonstrate the effectiveness of our proposed SDETR and COCO SOD dataset, we compare it with other state-of-the-art methods. As shown in Ta-

Method	DUTS[3]	ECSSD[11]	DUT[12]	HKU-IS[13]	PASCAL-S[14]
	$S_\alpha \uparrow F_\beta \uparrow E_\xi \uparrow M \downarrow$	$S_\alpha \uparrow F_\beta \uparrow E_\xi \uparrow M \downarrow$	$S_\alpha \uparrow F_\beta \uparrow E_\xi \uparrow M \downarrow$	$S_\alpha \uparrow F_\beta \uparrow E_\xi \uparrow M \downarrow$	$S_\alpha \uparrow F_\beta \uparrow E_\xi \uparrow M \downarrow$
CPD[15]	.869 .821 .898 .043	.913 .909 .937 .040	.825 .742 .847 .056	.906 .892 .938 .034	.848 .819 .882 .071
SCRN[16]	.885 .833 .900 .040	.920 .910 .933 .041	.837 .749 .847 .056	.916 .894 .935 .034	.869 .833 .892 .063
PoolNet[17]	.887 .840 .910 .037	.919 .913 .938 .038	.831 .748 .848 .054	.919 .903 .945 .030	.865 .835 .896 .065
BASNet[1]	.876 .823 .896 .048	.910 .913 .938 .040	.836 .767 .865 .057	.909 .903 .943 .032	.838 .818 .879 .076
EGNet[18]	.878 .824 .898 .043	.914 .906 .933 .043	.840 .755 .855 .054	.917 .900 .943 .031	.852 .823 .881 .074
F3Net[5]	.888 .852 .920 .035	.919 .921 .943 .036	.839 .766 .864 .053	.917 .910 .952 .028	.861 .835 .898 .062
ITSD[19]	.886 .841 .917 .039	.920 .916 .943 .037	.842 .767 .867 .056	.921 .906 .950 .030	.860 .830 .894 .066
SDETR	.898 .865 .933 .030	.933 .933 .958 .028	.855 .791 .883 .044	.921 .918 **.960** .026	.866 .851 .906 .057
SDETR+	**.903 .873 .938 .028**	**.937 .940 .964 .025**	**.865 .811 .902 .044**	**.922 .918** .958 **.026**	**.869 .855 .911 .055**

Table 1. Performance comparison with benchmark RGB salient object detection models. SDETR+ suggest the results with COCO SOD pretrain.

Layers	Model Size	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$
$[2, 2, 2, 2]$	13.1	.871	.825	.906	.0448
$[3, 4, 6, 3]$	24.2	.889	.851	.923	.0357
$[3, 4, 18, 3]$	44.0	.898	**.873**	**.938**	**.0295**
$[3, 8, 27, 3]$	60.8	**.901**	.867	.936	.0306

Table 2. Ablation study of Transformer encoder on DUTE[3].

Decoder	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$
PSP[21]	**.901**	.867	.935	.030
ASPP[22]	.900	.865	.935	.031
Ours	.898	**.873**	**.938**	**.029**

Table 3. Ablation study of SOD decoder on DUTE[3].

ble 1, our SDETR achieves the best performance across five datasets with respect to four metrics, which demonstrates the effectiveness of our predict-refine SDETR.

We also report the SDETR results with and w/o COCO SOD pretrain. SDETR+ denotes the results fine-tuned from COCO SOD pretrain. The results demonstrate that pretraining on COCO SOD improves SDETR performance on all datasets except HKU-IS datasets, demonstrating the PVT backbone is more advantageous than CNN in extracting the visual feature representations for pixel-level dense prediction of SOD.

Qualitative comparison. To give an intuitive understanding of the effectiveness of our method, we visualize sample results of SDETR and other sota methods on DUTS-TE datasets in Figure 4.

The comparative visualizations in Figure 4 also show that our PVT-based network can effectively integrate global understanding and reduce the structural loss for better salient detection, even under complex scenarios, such as complex backgrounds(Row 1 and 2), object occlusion(Row 3), multiple objects(Row 4). It leads to a balance between local texture details and salient structural integrity, benefiting from the global receptive field provided by the global attention mechanism of PVT backbone.

Method	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$M \downarrow$
Ours	**.903**	**.973**	**.938**	**.028**
w/o Refine Module	.901	.871	.935	0.29

Table 4. Ablation study of Refine Module on DUTE[3]

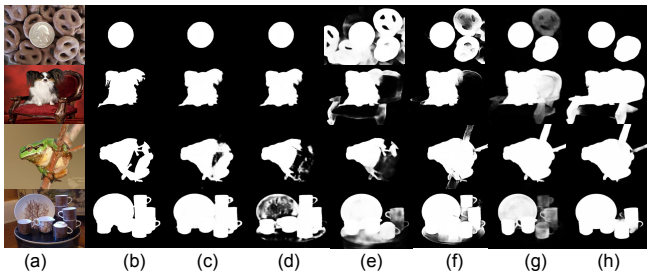

Fig. 4. Comparison results with state of the art RGB salient object detection models and our SDETR network model: (a)image, (b)GT, (c)Ours with COCO SOD pretrain, (d) Ours w/o COCO SOD pretrain, (e)F3net[5], (f)ITSD[19], (g)U2net[4], (h)BASNet[1].

We also observed that with our proposed COCO SOD as pre-train data, our model largely improves its ability to capture the saliency information under complex scenes in row.

6. CONCLUSION

In this paper, we propose a novel predict-refine framework named SDETR. We propose to apply a transformer-based network for coarse saliency map prediction. The coarse saliency map and error map are further fed into the refine module to generate detailed predictions. We also propose a pretrain dataset COCO SOD that can be widely used for future SOD tasks. Experimental results on five datasets demonstrate that SDETR outperforms previous methods under four evaluation metrics and COCO SOD pretrain can largely boost SDETR performance on four evaluation datasets.

7. REFERENCES

[1] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.

[2] Yuxin Mao, Jing Zhang, Zhexiong Wan, and Yuchao Dai *et al*, "Transformer transforms salient object detection and camouflaged object detection," *arXiv preprint arXiv:2104.10127*, 2021.

[3] Lijun Wang, Huchuan Lu, and Yifan Wang *et al*, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 136–145.

[4] Xuebin Qin, Zichen Zhang, and Chenyang Huang *et al*, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, pp. 107404, 2020.

[5] Jun Wei, Shuhui Wang, and Qingming Huang, "F^3net: Fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 12321–12328.

[6] Sixiao Zheng, Jiachen Lu, and Hengshuang Zhao *et al*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6881–6890.

[7] Wenhai Wang, Enze Xie, Xiang Li, and Deng-Ping Fan *et al*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," *arXiv preprint arXiv:2102.12122*, 2021.

[8] Ze Liu, Yutong Lin, and Yue Cao *et al*, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[9] Xiangxiang Chu, Zhi Tian, and Yuqing Wang *et al*, "Twins: Revisiting spatial attention design in vision transformers," *arXiv preprint arXiv:2104.13840*, 2021.

[10] Tsung-Yi Lin, Michael Maire, and Serge Belongie *et al*, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[11] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 1155–1162.

[12] Chuan Yang, Lihe Zhang, and Huchuan Lu *et al*, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166–3173.

[13] Guanbin Li and Yizhou Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5455–5463.

[14] Yin Li, Xiaodi Hou, and Christof Koch *et al*, "The secrets of salient object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 280–287.

[15] Zhe Wu, Li Su, and Qingming Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3907–3916.

[16] Zhe Wu, Li Su, and Qingming Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7264–7273.

[17] Jiang-Jiang Liu and Qibin Hou *et al*, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3917–3926.

[18] Jia-Xing Zhao and Jiang-Jiang Liu *et al*, "Egnet: Edge guidance network for salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8779–8788.

[19] Huajun Zhou, Xiaohua Xie, and Jian-Huang Lai *et al*, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9141–9150.

[20] Enze Xie and Wenhai Wang *et al*, "Segformer: Simple and efficient design for semantic segmentation with transformers," *arXiv preprint arXiv:2105.15203*, 2021.

[21] Hengshuang Zhao, Jianping Shi, and Xiaojuan Qi *et al*, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[22] Liang-Chieh Chen, George Papandreou, and Iasonas Kokkinos *et al*, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.