TRACER: Extreme Attention guided Salient Object Tracing Network

Min Seok Lee, Wooseok Shin, and Sung Won Han

Artificial Intelligence and Data Analytics Lab, School of Industrial and Management Engineering, Korea University

[**Code**] github.com/Karel911/TRACER
[**Arxiv**] arxiv.org/abs/2112.07380

AAAI-22

Introduction

In existing studies on salient object detection, performance gain and computational efficiency cannot be achieved, which has motivated us to study the inefficiencies in existing encoder-decoder structures to avoid this trade-off.

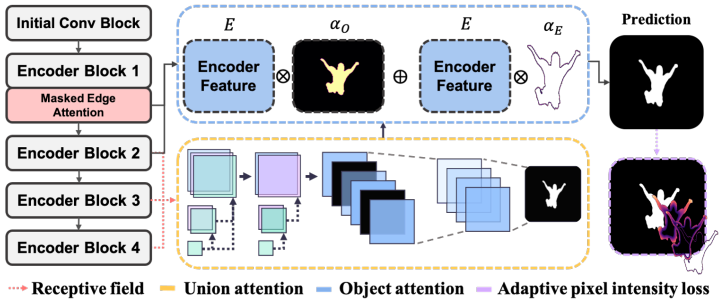
We propose TRACER which excludes multi-decoder structures and minimizes the learning parameters usage by employing attention guided tracing modules (ATMs).

Quantitative Results

Models	Size	#Params	GFLOPs	MPE	DUTS-TE			DUT-O			HKU-IS			ECSSD		
					MAE	S_m	FPS	MAE	S_m	FPS	MAE	S_m	FPS	MAE	S_m	FPS
SCRN	352	25.25M	30.18	10.04m	.040	.885	41.29	.056	.837	41.52	.033	.916	44.03	.037	.927	47.81
F3Net	352	25.54M	32.86	7.24m	.035	.888	60.51	.053	.838	63.22	.028	.917	78.63	.033	.924	76.37
LDF	352	25.15M	31.02	7.05m	.034	**.892**	64.41	.052	.839	67.00	.028	**.919**	82.89	.034	**.925**	79.92
TR-R	352	25.28M	25.94	3.73m	.035	.890	145.48	.050	.845	154.38	.028	**.919**	154.08	.033	**.925**	140.21
TE2	352	11.09M	5.20	**2.46m**	**.030**	.891	**242.92**	**.047**	**.846**	**267.25**	**.027**	.918	**260.35**	**.031**	.924	**231.31**

TE2 performed at least 2.9x to 4.1x faster than the existing methods on training each epoch and at least 3.8x to 6.4x faster on inference times. TR-R occupied approximately 12.9% of of total GFLOPs at the decoder structures. In contrast, the multi-decoder frameworks occupied 32.6% (SCRN), 38.2% (F3Net), and 34.6% (LDF) of total GFLOPs, respectively.

TRACER



* **Masked edge attention** - Existing methods cannot leverage the explicit edges in the feature extraction phases.

$$X_H = FFT^{-1}(f_r^H(FFT(X))) \quad (1)$$

$$X_E = X + \mathcal{RFB}(X_H) \quad (2)$$

* **Union attention** - To improve performance and network efficiency, the focus should be on determining which representations over multiple levels are important in the multi-level aggregation.

$$E_2^{'} = E_2 \otimes f(Up(E_3)) \otimes f(Up(Up(E_4))),$$
$$E_3^{''} = f(cat[E_3 \otimes f(Up(E_4)),\ f(Up(E_4))]), \quad (3)$$
$$E_2^{''} = f(Up(E_3^{''}))$$

$$\alpha_c = \sigma\left(\frac{\exp(\mathcal{F}_q(\widetilde{X})(\mathcal{F}_k(\widetilde{X}))^\top)}{\sum\exp(\mathcal{F}_q(\widetilde{X})(\mathcal{F}_k(\widetilde{X}))^\top)}\mathcal{F}_v(\widetilde{X})\right) \quad (4)$$

$$\widetilde{X}_c = X_c \otimes mask \begin{cases} mask = 1, & \text{if } \alpha_c > F^{-1}(\gamma) \\ mask = 0, & \text{otherwise} \end{cases} \quad (5)$$

$$D_0 = \frac{\exp(\mathcal{G}_q(\widetilde{X}_c)(\mathcal{G}_k(\widetilde{X}_c))^\top)}{\sum\exp(\mathcal{G}_q(\widetilde{X}_c)(\mathcal{G}_k(\widetilde{X}_c))^\top)}\mathcal{G}_v(\widetilde{X}_c) + \mathcal{G}_v(\widetilde{X}_c) \quad (6)$$

* **Object attention** - In contrast to the existing studies, we maintain a decoder feature as a single channel for decoder efficiency.

$$\alpha_E = \begin{cases} 0, & \text{if } (-\sigma(x_{ij}) + 1) > d \\ -\sigma(x_{ij}) + 1, & \text{otherwise} \end{cases} \quad (7)$$

$$D_{i+1} = \mathcal{RFB}((\alpha_O \otimes E_{2-i}) + (\alpha_E \otimes E_{2-i})) \quad (8)$$

* **Adaptive pixel intensity loss** - The BCE and IoU cause a class discrepancy between the foreground and background when all pixels are considered equally.

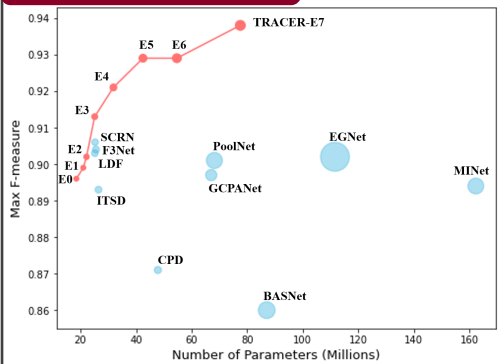
$$\omega_{ij} = (1-\lambda)\sum_{k\in K}\left|\frac{\sum_{h,w\in A_{ij}} y_{hw}^k}{\sum_{h,w\in A_{ij}} 1} - y_{ij}\right| y_{ij} \quad (9)$$

$$\mathcal{L}_{BCE}^a = -\frac{\sum_i^H\sum_j^W(1+\omega_{ij})\sum_{c=0}^1(y_c log(\hat{y}_c)+(1-y_c)log(1-\hat{y}_c))}{\sum_i^H\sum_j^W(1.5+\omega_{ij})} \quad (10)$$

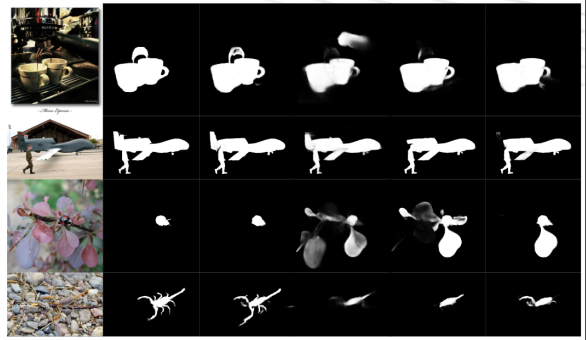
To improve the network for equivariance learning and to reduce the divergence discrepancy, we additionally measure the L1 distance for the final loss.

$$\mathcal{L}_{API}(y,\hat{y}) = \mathcal{L}_{BCE}^a(y,\hat{y}) + \mathcal{L}_{IoU}^a(y,\hat{y}) + \mathcal{L}_{L1}^a(y,\hat{y}) \quad (11)$$

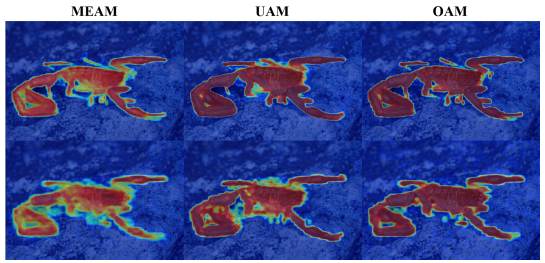
Qualitative Results



Comparison with model parameters and DUTS-TE MaxF scores. TRACER achieves outstanding performance and network efficiency compared to previous methods. Circle size indicates model GFLOPs.



The existing methods could not obtain the complete detailed region. For the detection of small-scale objects, TRACER could extract the salient objects precisely; in contrast, the previous methods identified objects that included the background or only a few areas.



* Heatmap visualization and comparison of results corresponding to the application of ATMs for the module explainability.

* When the ATMs are applied, the network not only discriminates the fine edges and redundant regions but also shows robustness against noise

Conclusion

* We studied the inefficiencies in the existing encoder-decoder structure to improve the SOD performance along with network efficiency.
* TRACER improves the performance and computational efficiency by employing ATMs and API loss in comparison to the three existing methods on the four bench-mark datasets.