示踪剂：极度注意力引导突出对象跟踪网络

张进，戴凌慧，滕炳杰，安思语

TRANCER: Extreme Attention Guided Salient Object Tracing Network

Zhang Jin, Dai Ling Hui, Ten Bing Jie, An Si Yu

Absrtact: Existing studies on salient object detection (SOD) focus on extracting distinct objects with edge information and aggregating multi-level features to improve SOD performance. To achieve satisfactory performance, the methods employ refined edge information and low multi-level discrepancy. However, both performance gain and computational efficiency cannot be attained, which has motivated us to study the inefficiencies in existing encoder-decoder structures to avoid this trade-off. We propose TRACER, which detects salient objects with explicit edges by incorporating attention guided tracing modules. We employ a masked edge attention module at the end of the first encoder using a fast Fourier transform to propagate the refined edge information to the downstream feature extraction. In the multi-level aggregation phase, the union attention module identifies the complementary channel and important spatial information. To improve the decoder performance and computational efficiency, we minimize the decoder block usage with object attention module. This module extracts undetected objects and edge information from refined channels and spatial representations. Subsequently, we propose an adaptive pixel intensity loss function to deal with the relatively important pixels unlike conventional loss functions which treat all pixels equally.

Keywords: SOD；TRACER

摘要：现有的显著目标检测（SOD）研究主要是提取具有边缘信息的不同目标并聚合多级特征来提高 SOD 性能。为了获得令人满意的性能，这些方法采用精细的边缘信息和低多级差异。然而，性能提升和计算效率都无法实现，这促使我们研究现有编码器-解码器结构中的低效率，以避免这种权衡。我们提出了 TRANCER，它通过结合注意力引导跟踪模块来检测具有明确边缘的突出物体。我们在第一个编码器的末端使用了屏蔽边缘注意力模块，适用快速傅里叶变换将

细化的边缘信息传播到下游特征提取。在多级聚合阶段，联合注意力模块识别互补通道和重要空间信息。为了提高解码器的性能和计算效率，我们使用对象注意力模块最大限度减少解码快的适用。该模块从细化的通道和空间表示中提取未检测到的对象和边缘信息。随后，我们提出一种自适应像素强度损失函数来处理相对重要的像素，这与平等对待所有像素的传统损失函数不同。与 3 种现有的方法的比较表明，TRACER 在三个基准数据集上实现了较先进的性能。

关键词：显著物体目标检测 示踪剂

为了提高显著目标检测（SOD）的性能，现有方法可以分为两种，分别是改进边缘表示的方法和减少多级聚合过程中差异的方法。但是由于在显著物体检测的过程中，存在目标背景复杂，边缘细节不够突出，算法复杂程度较高，计算效率低等等的问题，目前所提出的显著物体检测始终还不能很好地接近我们人工判断的理想的显著物体检测目标理想图。文献[1]提出了一个简单而强大的深度网络架构，U2NET，用于显著目标检测（SOD），其架构是一个两级嵌套的 U 结构。由于其中混合了不同大小的感受野，所以能够从不同尺度捕获更多的上下文信息；并且由于其中的 U 结构中使用了池化操作，增加了整个架构的深度，而不会显著增加计算成本，这种 U2NET 结构能够从头开始训练深度网络，而无需使用图像分类任务的主干。文献[2]提出了一种自适应和专注深度蒸馏器，A2dele，用于高效的 RGB-D 突出物体检测。由于现有的最先进的 RGB-D 突出目标检测方法依赖于双流架构探索 RGB-D 数据，其中需要独立的子网来处理深度数据，不能避免会产生额外的计算成本和内存消耗，并且在测试过程中使用深度数据可能会阻碍 RGB-D 显著性检测的实际应用，通过使用 A2dele 来探索使用网络预测和

注意力作为将深度认知从深度流转移到 RGB 流的桥梁的方式来提高显著物体检测的性能。但是这种方法在分流上所耗费的时间成本比较高，大大提升了训练的成本，并且所得到的结果与现有的先进方法对比所达到的性能效果并不高。

本文使用的是一种称为示踪剂（TRACER）的极度注意力引导突出对象跟踪网络。为了解决现有方法的低效率问题，我们在浅层编码器、多级聚合过程和解码器中应用了三个注意力引导模块（即屏蔽边缘、联合和对象衰减模块）。屏蔽边缘衰减模块使用快速 傅里叶变换增强低级表征中的边缘特征，并将边缘优化表示传播到下一个编码器。联合注意力模块聚合多级编码器输出，以减少分布中的差异。随后，该模块确定聚合门控通道和空间表示中和解码器输出，以识别突出的物体。为了处理像素的相对重要性，我们提出了一个自适应像素强度损失函数。通过聚合多核聚合来聚合目标像素周围的相邻像素，并排除边缘外的权重。当目标像素由精细或明确的边缘组成时，为其分配的强度高于其他像素，以此来提高整个系统的性能。

后面章节的内容安排：算法框架以及算法模块的描述，其中包含：1.体系结构概述、2.注意力引导追踪模块、3.自适应像素强度损失函数；实验结果分析以及小组成员分工。

一、算法体系结构以及其描述

1. 体系结构概述

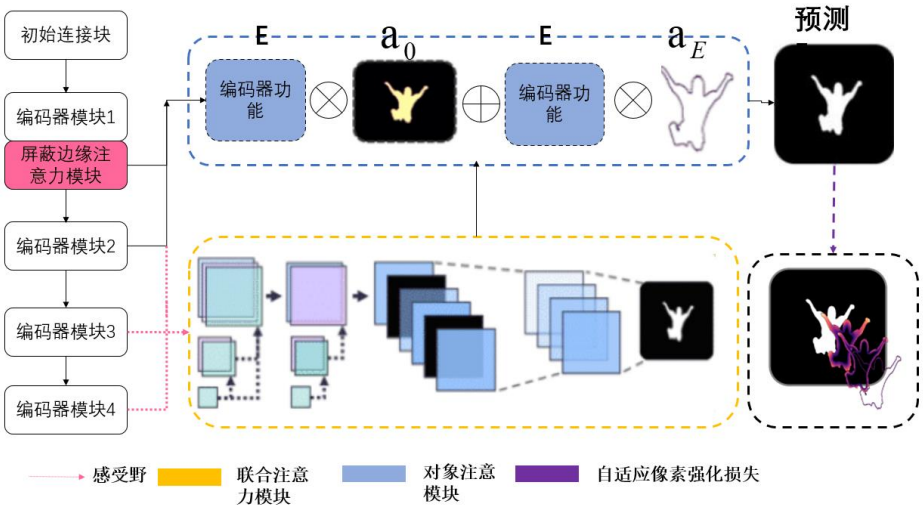


图 1：体系结构

由于现有的主干编码器在特征提取性能和内存效率方面存在漏洞，因此需要替代主干。因此，我们使用 Efficient(Tan 和 Le 2019)作为主干编码器，并将现有的七个块合并为四个块，其中输出分辨率被移位，除了初始卷积块。首先将屏蔽边缘注意力模块初始应用于具有足够边界感表示的第一个编码器块输出为 E_1，目的是增强边缘信息并提高内存效率。在解码器中实现了联合注意力模块，分别聚合了多层特征并合并了编码器和解码器的输出，在联合注意力模块中，基于多核的接受场块得到的三个编码器的输出 E_2、E_3、E_4，以 E_2 的尺度进行多层集合。

联合注意力模块强调更清晰的渠道和空间信息，对象注意模块提取具有互补边缘信息的明显目标，并利用这些补充信息减小浅层编码器与图像的差异。此外，对象注意力模块由深度卷积块组成，尽量减少学习参数的数量，提高计算效率。最后，TRACER 生成四个深度监督图（DS_i），它们是并集联合（DS_0）对象注意力模块（DS_1 和 DS_2）和 DS 映射集合（DS_e）的输出。

2. 注意引导跟踪模块

检测具有边缘的不同物体对于提高 SOD 性能至关重要。使用卷积模块（为

提高计算效率），通过注意力引导的突出对象跟踪模块（ATM）跟踪对象和边缘以提高性能。

2.1 屏蔽边缘注意力模块

为了跟踪边缘信息，本文提出了屏蔽边缘注意力模块（MEAM），通过使用快速傅里叶变换（FFT）提取显式边界，建立第一个编码器的输出边界。现有的方法使用边缘信息，但是它们无法在特征提取阶段利用显式边缘，由于这些方法需要深度编码器的输出才能获得不同的边缘。因此，我们使用 FFT 仅从第一个编码器表示中提取显式边缘。同时使用傅里叶变换及其逆变换，将第一个编码器表示为高频率和低频率，如下所示：

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

其中 X 表示输入特征，f_r^H 是一个高通滤波器，它消除了除半径 r 以外的所有频率，为了判别显式边缘，我们利用了高通滤波器获得的高频，这些滤波器具有足够的边界信息。此外，当 X_H 包含了背景噪声，我们采用接受场运算 RFB 来消除噪声。通过感受野操作生成显式边缘，最后计算精细边缘，明确边缘损失。

2.2 联合注意力模块

联合注意力模块的（UAM）的目的是聚合多层次的功能，并检测更重要的上下文通过和空间表示。在这里，$f(.)$ 和 $cat(.)$ 分别表示卷积运算和通道特性链接。每个编码器输出分别聚合到 32、64 和 128 个通道，积分如下：

$$\begin{aligned} E_2' &= E_2 \otimes f(U_p(E_3)) \otimes f(U_p(U_p(E_4))), \\ E_3'' &= f(cat[E_3 \otimes f(U_p(E_4)), f(U_p(E_4))], \\ E_2'' &= f(U_p(E_3'')) \end{aligned} \tag{2}$$

得到的一个聚集表示，通过 $X = f(cat[E_2', E_2'']) \in \mathbb{R}^{(32+64+128)\times H_2 \times W_2}$ 聚合后，仍然是上下文信息是相对显著的信道和空间特征。然而，现有的研究已经将通道和空间注意力模块独立于解码器和感受野块，尽管这两个空间的依赖性。因此，我们同时要强调空间信息的基础上获得互补的信息通道的上下文。

$$\alpha_c = \sigma\left(\frac{\exp(\mathcal{F}_q(\widetilde{X})(\mathcal{F}_k(\widetilde{X}))^\top)}{\sum\exp(\mathcal{F}_q(\widetilde{X})(\mathcal{F}_k(\widetilde{X}))^\top)}\mathcal{F}_v(\widetilde{X})\right) \tag{3}$$

Xe∈Rc×1×1 是信道分组表示，F(·)表示使用 1×1 核大小的卷积运算。有效通道 αc∈Rc×1×1，得到上下文信息。为了细化聚合表示 X，我们应用置信通道权重如下：xc=(X⊗αc)+X。随后，根据 αc 的分布和置信比 γ 保留信任通道，如下所示：

$$\widetilde{X}_c = X_c \otimes mask \quad \begin{cases} mask = 1, & \text{if } \alpha_c > F^{-1}(\gamma) \\ mask = 0, & \text{otherwise} \end{cases} \tag{4}$$

这里，$F^{-1}(\gamma)$ 表示 αc 的 γ 分位数。我们排除了分布 αc 下部尾部的 γ 区域。然后，在空间上计算精炼的输入 Xec 以区分显著对象并产生第一解码器表示，$D_0 \in \mathbb{R}^{1\times H_2 \times W_2}$，如等式 5 所示。

$$D_0 = \frac{\exp(\mathcal{G}_q(\widetilde{X}_c)(\mathcal{G}_k(\widetilde{X}_c))^\top)}{\sum\exp(\mathcal{G}_q(\widetilde{X}_c)(\mathcal{G}_k(\widetilde{X}_c))^\top)}\mathcal{G}_v(\widetilde{X}_c) + \mathcal{G}_v(\widetilde{X}_c) \tag{5}$$

2.3 对象注意模块

为了减少使用最小参数的编码器和解码器表示之间的分布差异，我们组织了一个对象注意模块（OAM）作为解码器。与现有的研究（Chen et al.2018；Zhao et al.2019）相比，我们将 D 作为解码器效率的单一通道，并且 OAM 跟踪来自每个解码器表示 $D_i \in \mathbb{R}^{1\times H\times W}$ 的对象边缘和互补边缘。同时进行细化突出物操作，由于不能检测到整个对象与明确的边缘区域，因此，我们生成一个互补的边缘权重

去覆盖未检测到的区域。对于 D 中的每个像素我们反转检测到的区域，并消除与缺失区域检测的去噪比相对应的背景噪声：

$$\alpha_E = \begin{cases} 0, & \text{if } (-\sigma(x_{ij}) + 1) > d \\ -\sigma(x_{ij}) + 1, & \text{otherwise} \end{cases} \tag{6}$$

我们结合编码器输出和解码器特征：

$$D_{i+1} = \mathcal{RFB}((\alpha_O \otimes E_{2-i}) + (\alpha_E \otimes E_{2-i})) \tag{7}$$

为了减少差异，我们利用感受野操作 RFB（·）和上采样 Di+1 产生 DSi+1。

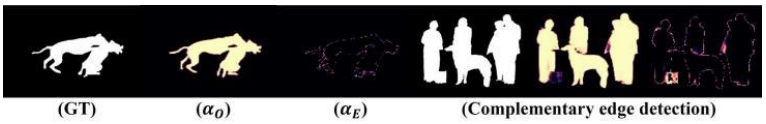


图 2：物体互补边缘检测

3. 自适应像素强度损失函数

对于损失函数，我们结合二进制交叉熵（BCE），LoU，和 L1 损失函数来减少对象和背景之间的差异。虽然二进制交叉熵和 LoU 在全球范围内采用的损失函数，这些功能会导致类之间的差异的前景和背景时，所有的像素被认为是平等的。与显著对象的背景和中心的像素相比，与精细或显式边缘相邻的像素需要更多的关注。因此，我们提出自适应像素强度（API）损失，将像素强度应用于每个像素，如下所示：

$$\omega_{ij} = (1 - \lambda) \sum_{k \in K} \left| \frac{\sum_{h,w \in A_{ij}} y_{hw}^{k}}{\sum_{h,w \in A_{ij}} 1} - y_{ij} \right| y_{ij} \tag{8}$$

我们通过使用多个内核大小 K 并排除边缘之外的权重，聚合目标像素周围的相邻像素。如果目标像素由细边缘组成，则采用多核聚合来为目标像素分配比其他像素更多的权重。如下图所示：

图 3：对应内核大小 K 的像素强度 w 可视化

像素强度用于自适应损失函数损耗。通过使用像素强度和自适应损失函数使得网络更多地关注与显式或细边相关的局部结构，而不像 BCE 损失。

$$\mathcal{L}^{a}_{BCE} = -\frac{\sum_{i}^{H}\sum_{j}^{W}(1+\omega_{ij})\sum_{c=0}^{1}(y_c log(\hat{y}_c)+(1-y_c)log(1-\hat{y}_c))}{\sum_{i}^{H}\sum_{j}^{W}(1.5+\omega_{ij})} \tag{9}$$

与此相反，自适应的 LoU 损失优化的全局结构的基础上的密集特征对应 ω。如（10）所示，与原始的 loU 损失相比，对于密集区域高度关联的像素进行了区分和强调。

$$\mathcal{L}^{a}_{IoU} = 1 - \frac{\sum_{i}^{H}\sum_{j}^{W}(y_{ij}\hat{y}_{ij})(1+\omega_{ij})}{\sum_{i}^{H}\sum_{j}^{W}(y_{ij}+\hat{y}_{ij}-y_{ij}\hat{y}_{ij})(1+\omega_{ij})} \tag{10}$$

此外，为了进一步提高网络的等方差学习，以减少分歧的差异，我们测量的 L1 距离，这使网络学习强劲对嘈杂的标签（Ghosh，Kumar 和 Sastry 2017 年；Wang 等人 2020 年）。L1 损失同样处理所有像素；因此，我们将像素强度 ω 应用到 L1 损失以区分相对显著的像素，并排除邻近明确或精细边缘的噪声像素，如下所示：

$$\mathcal{L}_{L1}^{a}=\frac{\sum_{i}^{H}\sum_{j}^{W}|y_{ij}-\hat{y}_{ij}|(1+\omega_{ij})}{H\times W\sum_{i}^{H}\sum_{j}^{W}\omega_{ij}} \tag{11}$$

为了合并上述局部和全局结构强度，我们将 API 损失函数组合为:

$$\mathcal{L}_{API}(y,\hat{y})=\mathcal{L}_{BCE}^{a}(y,\hat{y})+\mathcal{L}_{IoU}^{a}(y,\hat{y})+\mathcal{L}_{L1}^{a}(y,\hat{y}) \tag{12}$$

在组合损失函数的基础上，利用地面真值 G、DSi∈{0，1，2}三种深度监督，三种监督 DSE 的集合，以及由 MEAM 获得的显式边缘 E，对最终损失进行了优化，结果如下：

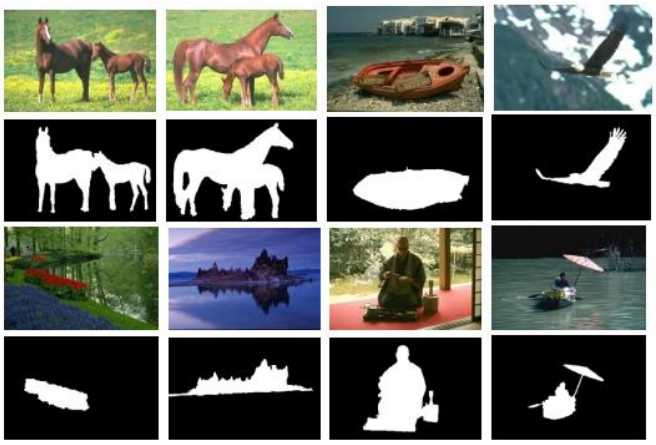
$$\mathcal{L}=\sum_{i}\mathcal{L}_{API}(G,DS_i)+\mathcal{L}_{API}(E,\hat{E}) \tag{13}$$

二、实验结果分析

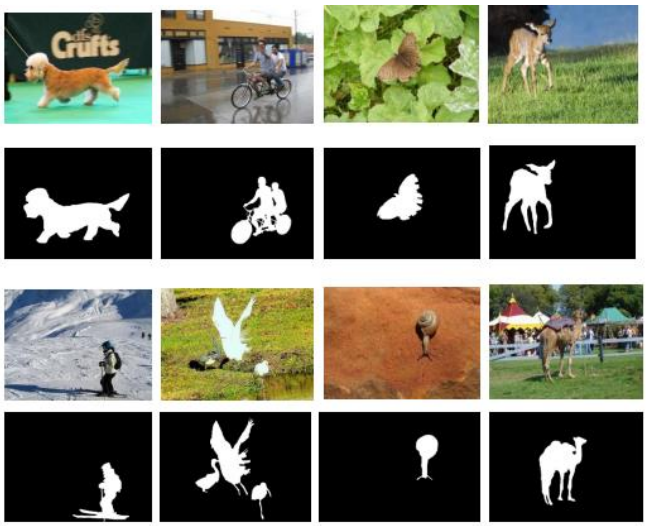
1. 数据

我们本次实验对三个基准数据集进行了评估：ECSSD，DUTS-TE，DUT-OMRON。以下是三个数据集的部分展示：

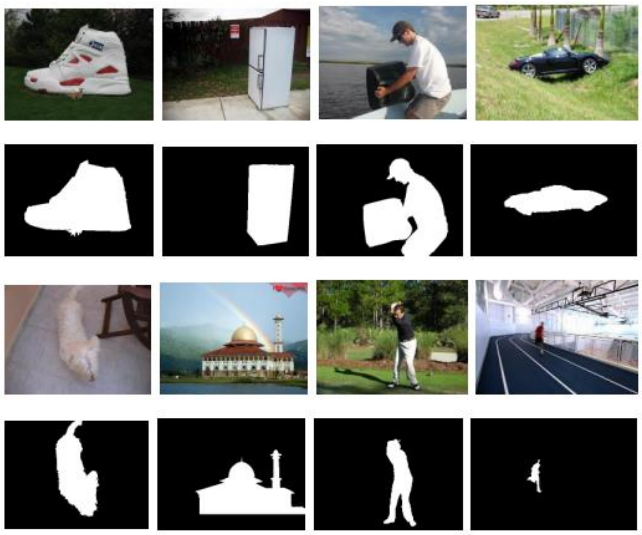
- ECSSD



- DUTS-TE



- DUT-OMRON



其中 ECSSD （Yan et al. 2013）包含 1，000 个结构复杂且语义有意义的场景。DUTS（王等. 2017）是 SOD 最大的基准数据集。它包含 10，553 张训练图像和 5，019 张测试图像。 DUT-OMRON（Yang 等人。 2013）有 5，168 张图像，其中包括一个或多个具有相对复杂背景的突出物体。由于在实际实验过程中，运行内存的原因，我们将其中的 DUTS-TE，DUT-OMRON 数据集进行了部分删减，最终所使用的数据集中：ECSSD 数据集 1，000 张，未删

减、DUTS-TE 数据集 687 张、DUT-OMRON 数据集 1，729 张。

2. 性能指标评估

在我们实验的性能评估阶段，我们总共选用了 8 个性能指标来评估我们本次实验的结果，同时将其与三个现有的算法进行对比，分别是：CVPR_depth、CVPR_rgb、U2NET。使用的 8 个性能指标分别是：最大的 F-measure（maxf）、平均阈值（fm）、平均绝对误差（MAE）、W-fmeasure、S-measure、E-measure，以及通过计算的精确率-召回率所绘制的 PR 曲线。其各个指标的说明如下：

- Maxf= max(1.3*prec.*rec./(0.3*prec+rec+eps))

  注：其中的prec指的是精确率，rec指的是召回率。

- F-measure = (2 * Precision * Recall) / (Precision + Recall)

- $MAE = \frac{1}{\mathrm{m}}\sum_{i=1}^{m}|y_i - y_i'|$

  注：其中的 y_i 指的是真实值，即实验过程中理想的mask， y_i' 是我们通过算法所预测的结果图。

- W-Fmeasure是综合评价指标F-measure的中的一个评价指标。

- S-measure是计算前景像素和真值的相似度，其计算方法如下：

  $$S_{\mathrm{m}} = \alpha * S_0 + (1-\alpha) * S_r$$

  $$S_0 = \frac{2 * E_{(pre)}}{E_{(pre)}^{\ 2} + 1 + \sigma + e}$$

  其中 S_r 表示按重心位置，然后切割成四个区域，对四个区域按像素占整张图的区域面积作为权重，计算四个区域SSIM的加权平均。最后实验计算需要对所有的sm做一个平均。

- E-measure用来评价预测结果中匹配到正确结果的百分比。

- PR曲线的组成主要是由精确率和召回率进行绘制的，其计算方法如下：

$$\Pr ecision = \frac{TP}{TP + FP}$$

$$\mathrm{Re}call = \frac{TP}{TP + FN}$$

其中的TP、FP、FN、TN的表示意义可以用下面的图来表示：

	预测	
真实	TP真正例	FN假负例
	FP假正例	TN真负例

TP-实例为正类，实际预测为正类；TP-实例为负类，实际预测为负类；FP-实例为负类，实际预测为正类；TN-实例为正类，实际预测为负类。

3. 实验设置

在整个实验的过程中，我们所使用的 python3.9，编程环境是 pycharm，在实现该算法的过程中所需要的包在说明文档 readme 中可查看。文中的算法以及对比的算法均可运行实现。在实验过程中，我们使用 DUTS-TR 数据集进行训练，并选用其它三个数据集 ECSSD，DUTS-TE，DUT-OMRON 进行本次实验的测试阶段。在实验中，我们的超参数设置：将其大小设置为 32，将周期数设置为 100. 我们使用亚当优化器学习率为5×10^{-5}，权重衰减为10^{-4}。我们观察到损失在五个时期后没有减少。

4. 实验结果展示及分析

4.1 定性分析结果

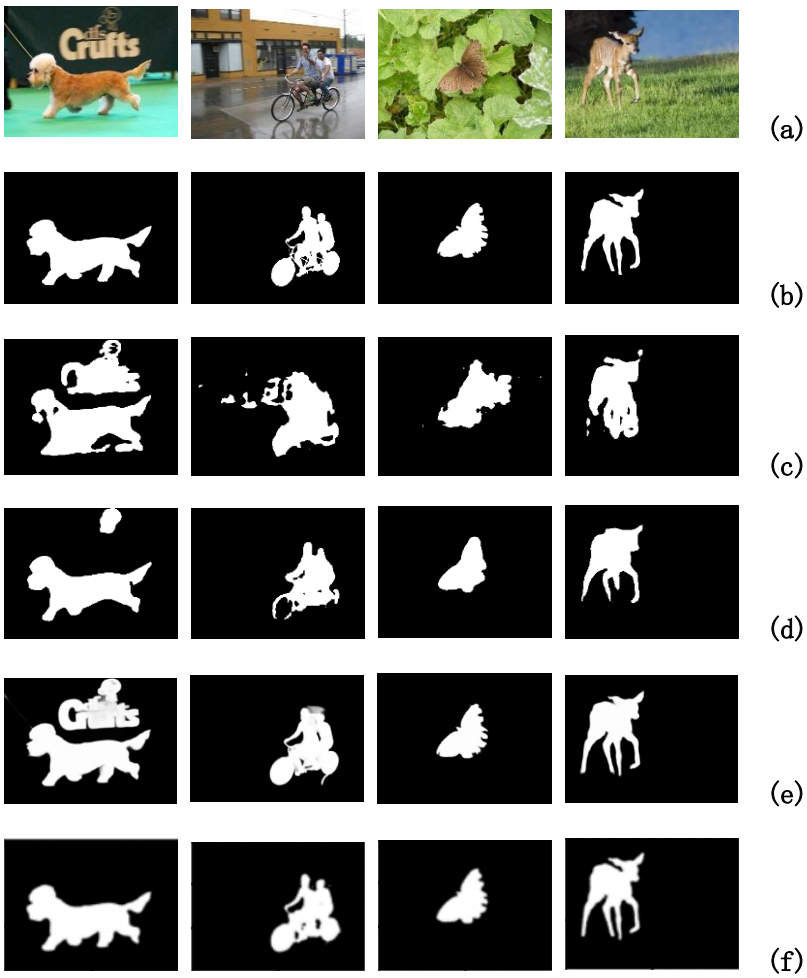


Figure1：DUTS-TE 数据集：(a)原图；(b)理想 maks；(c)CVPR-DEPTH；(d)CVPR-RGB; (e)U2NET；(f)ours

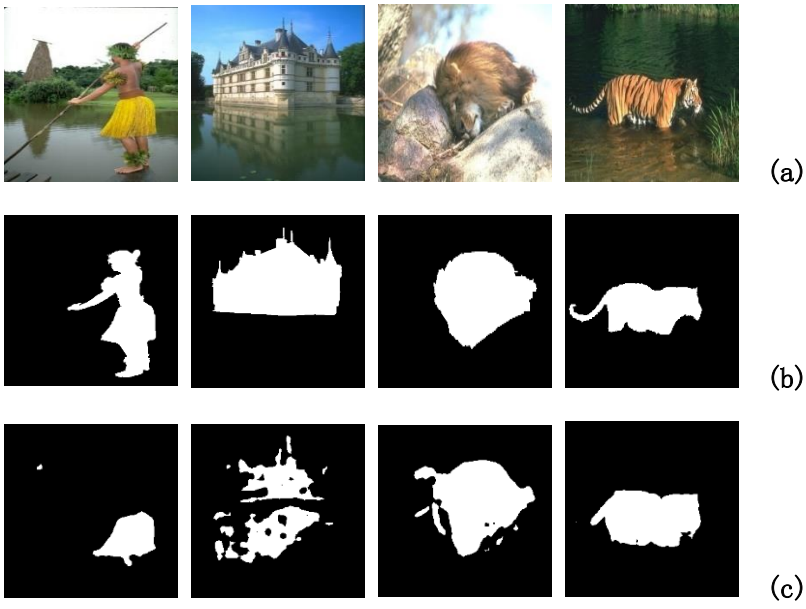
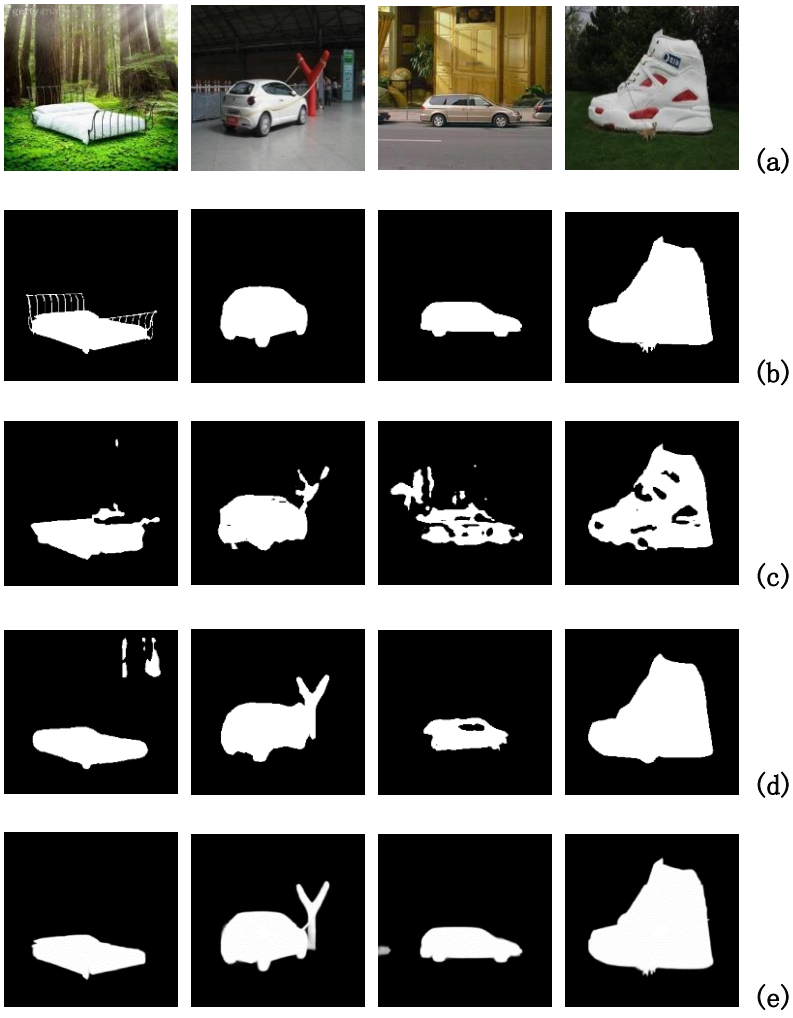


Figure2：ECSSD 数据集：(a)原图；(b)理想 maks；(c)CVPR-DEPTH；(d)CVPR-RGB; (e)U2NET；(f)ours



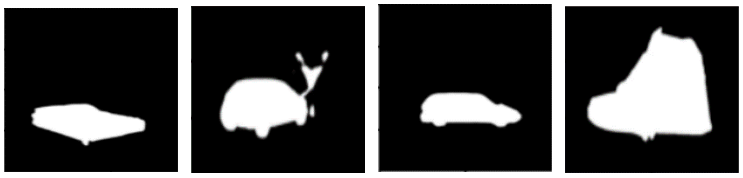

(f)

Figure3：DUT-OMRON 数据集：(a)原图；(b)理想 maks；(c)CVPR-DEPTH；(d)CVPR-RGB; (e)U2NET；(f)ours

由定性的实验结果可以看出，我们的算法结果和 U2NET 的算法结果在实现效果上是最好的，但是两种算法在实现显著性目标检测目标存在细边缘的时候，所检测的结果会忽略这一部分的突出目标，而将其视作是背景的部分。其次，在一些显著物体检测时候，会把一些颜色突出的部分当作目标检测出来，同时我们的算法在边缘细节的处理上还不是很好，不能很好的处理存在细边缘的显著性图片。相比之下，在显著性目标检测上，我们的算法所达到的准确性还是比较高的，都能准确地找到数据集中的显著性目标，但是在边缘细节的处理上，我们的算法所达到的效果没有 U2NET 算法达到的的效果好。我们分析其中的原因，是由于在我们的算法中，我们使用了屏蔽边缘注意力模块以及自适应像素强度损失函数对具有明显边缘信息的显著性目标的像素进行了权重划分，一定程度上弱化了边缘信息，但正是因为我们使用这样的方法，我们的算法性能相比于其它算法具有了明显的优势。

4.2 定量分析结果

Algorithm	ECSSD 1,000 images						DUTS-TE 687 images						DUT-OMRON 1,729 images					
	Maxf	Fm	MAE	wfm	Sm	EM	Maxf	Fm	MAE	wfm	Sm	EM	Maxf	Fm	MAE	wfm	Sm	EM
CVPR_depth	0.745	0.708	0.123	0.645	0.712	0.795	0.647	0.610	0.101	0.559	0.693	0.797	0.542	0.492	0.132	0.438	0.611	0.721
CVPR_rgb	0.901	0.878	0.058	0.847	0.861	0.898	0.829	0.801	0.048	0.771	0.834	0.889	0.705	0.656	0.071	0.614	0.705	0.756
U2NET	0.951	0.892	0.033	0.910	0.928	0.924	0.908	0.834	0.032	0.851	0.906	0.917	0.821	0.758	0.062	0.750	0.837	0.867
ours	**0.950**	**0.961**	**0.026**	**0.908**	**0.935**	**0.928**	**0.906**	**0.932**	**0.022**	**0.914**	**0.910**	**0.914**	**0.828**	**0.849**	**0.045**	**0.840**	**0.855**	**0.850**

表 1：基于 TRACER 的方法的定量分析结果

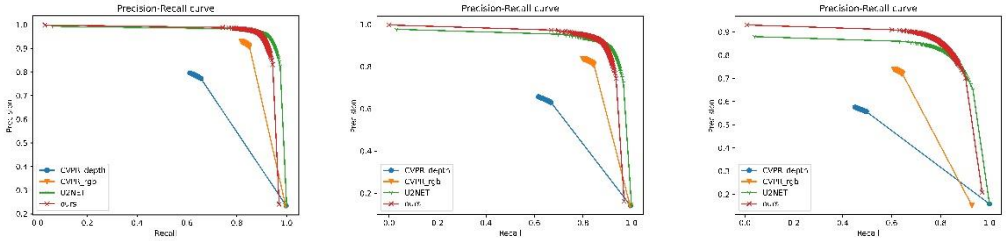


图 4：ECSSD、DUTS-TE、DUT-OMRON 数据集下的各算法的 PR 曲线

其中 Maxf、Fm、Wfm、Sα（Sm）、Eξ（EM）越大越好、MAE 越小越好。由定量指标的结果可知，我们基于 TRACER 算法所实现的显著性目标检测所能达到的效果对于于其他三个算法都具有明显的优势，但是在与 u2net 算法作比较的时候，我们的数据集在某些指标上的效果没有 u2net 算法的好，正是由于我们的 TRANCER 在算法中加入了屏蔽边缘模块和自适应像素强度损失函数模块，所以在边缘处理上的效果没有 u2net 算法所达到的效果好。并且 PR 曲线的结果来看，我们的算法还是具有明显优势的。总体来看，我们参考所实现的算法在现有的三个基准数据集上具有很好的竞争力。

5. 小组成员贡献

小组成员	张进	腾炳杰	戴凌慧	安思语
主要负责内容	PPT 汇报负责人、算法的改进和调整、数据整理。	Trancer 算法的实现和调整、PPT 部分内容撰写、运行说明文档撰写。	对比三个算法的实现、性能指标算法的实现、研究课程论文文档的撰写、算法流程框图梳理。	数据集的收集、PPT 的制作、汇总 PPT 的制作、论文算法步骤梳理。
贡献比	20%	25%	30%	25%

参考文献：

[1]Xuebin Qin,Zichen Zhang,Chenyang Huang,Masood Dehghan,Osmar R. Zaiane,Martin Jagersand. U 2 -Net: Going deeper with nested U-structure for salient object detection[J]. Pattern Recognition,2020,106(C).

[2]. Piao Y.,Rong Z.,Zhang M.,Ren W.,Lu H.. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2020.