基于深监督非线性融合的显著性物体检测

Yun Liu, Ming-Ming Cheng, Xin-Yu Zhang, Guang-Yu Nie, and Meng Wang

摘要—显著性物体检测的最新进展主要针对于如何有效地融合来自卷积神经网络（Convolutional Neural Network, CNN）的多尺度卷积特征。许多流行的方法均通过使用深监督来进行侧输出预测，这些预测被线性融合以进行最后的显著性预测。在这篇论文中，我们分别从理论和实验上证明了对侧输出预测进行线性融合不是最优的，这种融合方式仅对由深监督获得的侧输出信息进行了有限地利用。为了解决上述问题，我们提出了深监督非线性融合（Deeply-supervised Nonlinear Aggregation, DNA）来更好地利用不同侧输出之间的互补信息。它与现有方法相比有两点不同，首先，它融合的是侧输出特征而不是侧输出预测；其次，DNA 使用非线性变换而不是线性变换。实验表明，DNA 可以成功突破当前线性融合方法所带来的瓶颈。具体而言，所提出的显著性物体检测器，即一个改良过的具有 DNA 的 U-Net 架构，在不需要复杂的工程技巧的前提下，在各种数据集和多个评测指标上均超过当前最新的方法。

关键词—显著性物体检测，显著性检测，深监督非线性融合。

I. 引言

显著性物体检测（也称为显著性检测）旨在模拟人类视觉系统来检测自然图像中最明显且最吸引眼球的物体或区域 [1], [2], [3]。显著性物体检测的发展对很多视觉应用都有所帮助，包括图像检索 [4], [5]、视觉跟踪 [6], [7]、场景分类 [8]、内容感知的图像/视频处理 [9], [10]、缩略图生成 [11]、视频物体分割 [12], [13]、照片裁剪 [14] 和弱监督学习 [15], [16] 等。尽管已经有许多模型被提出 [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] 并且取得了重大突破，但是准确地检测静态图像中完整的显著性物体（尤其是在复杂场景中）仍然是一个有待解决的问题。

Manuscript received April 19, 2005; revised August 26, 2015. This research was supported in part by NSFC (NO. 61620106008, 61572264), in part by the national youth talent support program, in part by Tianjin Natural Science Foundation for Distinguished Young Scholars (NO. 17JCJQJC43700), and in part by Huawei Innovation Research Program.

Y. Liu, M.-M. Cheng, and X.-Y. Zhang are with Nankai University. M.-M. Cheng (cmm@nankai.edu.cn) is the corresponding author.

G.-Y. Nie is with Beijing Institute of Technology.

M. Wang is with Hefei University of Technology.

传统的显著性物体检测方法 [2], [30], [31] 通常设计手工制定的底层特征和启发式先验，这些方式很难表征语义对象和场景。显著性物体检测的最新研究主要基于卷积神经网络（Convolutional Neural Network, CNN）[32], [33], [34], [35], [36], [37]。一方面，通过逐渐增大的感受野和下采样尺度 [38]，CNN 可以自然地在每一层学习多尺度和多层次的特征表示。另一方面，由于图像内及图像间的物体/场景的尺度各不相同，所以显著性物体检测需要多尺度学习 [39], [40]。因此，当前最先进的显著性物体检测器 [18], [41], [42], [43], [20], [44], [45], [46], [47] 主要旨在设计复杂的网络体系结构来利用多尺度的 CNN 特征，即高层语义信息和与之互补的底层空间细节信息。

由于 U-Net [28]（或 FCN [48]）和 HED [29] 在多尺度学习中的优越性，许多领先的显著性物体检测器均在 U-Net 网络的基础上添加了深监督 [19], [42], [20], [21], [45], [49], [22], [50], [51]（如图 1(d) 所示）。我们注意到，这些网络首先使用侧输出来预测多尺度显著性图，然后通过逐像素卷积（即 1×1 卷积）等操作将生成的多尺度侧输出预测进行线性融合，以此来获得最终的显著性预测，从而可以有效地利用所有侧输出预测的优势。但是，我们在理论上和实验上均证明了这种线性融合侧输出预测的做法不是最优的，它对侧输出特征中的互补的多尺度信息的利用是有限的。这一点，我们将在第 III节中提供更详细的证明。

不同于线性融合侧输出预测，我们提出了一种非线性侧输出融合的方法。具体来说，我们将侧输出特征而不是侧输出预测进行拼接，然后使用非线性变换来预测显著性物体。与此同时，如图 1(e) 所示，我们还对侧输出特征加入了深监督，使其在训练阶段能更好地进行优化。用这种方式，所拼接的特征可以更好地利用多尺度侧输出特征。我们将所提出的方法命名为深监督非线性融合（Deeply-supervised Nonlinear Aggregation, DNA）。我们将 DNA 用于经简单修改后的 U-Net，在不需要复杂的工程技巧的前提下，所提出的网络即可取

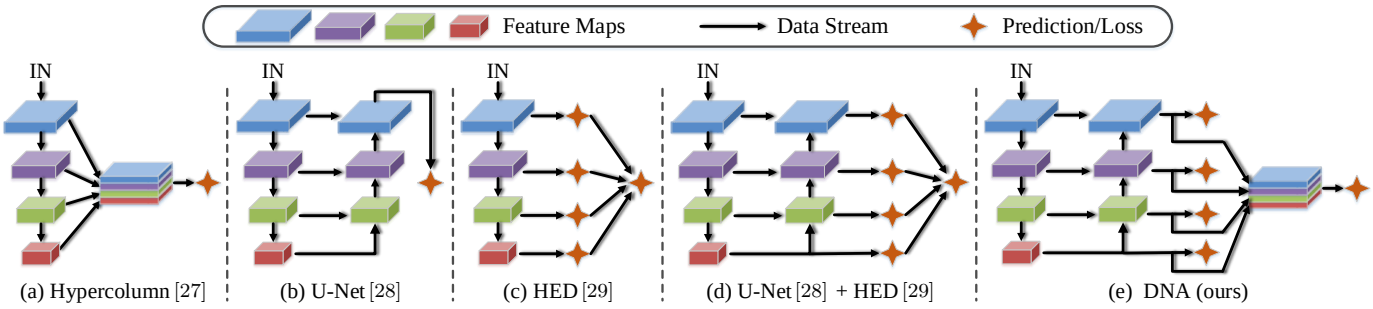

图 1. 不同的多尺度深度学习架构示意图。注意，虽然 (c)-(e) 均对侧输出使用了深监督，但是 (c) 和 (d) 线性融合了侧输出预测，而所提出的 DNA (e) 对侧输出特征采用了非线性融合。

得比当前最先进的显著性物体检测器都要好的效果，并且具有更少的参数和更快的检测速度。这里，我们的创新点有以下两点：

- 我们在理论和实验上分析了传统线性侧输出融合的局限性，这种局限性只能对多尺度侧输出信息进行有限地利用；
- 我们提出了用于侧输出特征的深监督非线性融合方法（DNA），其有效性已通过将其引入参数更少、速度更快的简单网络中而得到证明。

II. 相关工作

显著性物体检测因其广泛的应用范围和具有挑战性的场景而成为一个非常活跃的研究领域。早期的启发式显著性物体检测方法手工提取底层特征，然后使用机器学习模型对这些特征进行分类 [52], [53], [54], [55]，并利用一些启发式的显著性先验来确保其准确性，例如颜色对比 [1], [2]、中心优先 [56], [30] 和背景优先 [57], [58] 等。由于深度 CNN 在计算机视觉领域取得了巨大成功，基于 CNN 的方法也被用于显著性物体检测 [59], [60], [61], [62], [63]。基于区域的显著性物体检测 [64], [65], [66], [67], [68], [23], [24] 出现在早期的基于深度学习的显著性物体检测中。这种方法将每个图像小块视为进行显著性检测的基本处理单元。近年来，基于 CNN 的“图像-图像”的显著性物体检测方法 [18], [19], [41], [42], [43], [20], [44], [21], [45], [69], [46], [70], [71], [47], [72], [73], [49], [25], [26] 将显著性物体检测视为逐像素的回归任务并进行“图像-图像”的预测，在该领域占据了主导地位。因此，下面我们主要回顾一下基于 CNN 的“图像-图像”的显著性物体检测。

由于显著性物体检测需要高层的全局信息（存在于 CNN 的顶层）和底层的局部细节（存在于 CNN 的低层），因此如何有效地融合多层深度特征是主要的研究方向 [18], [41], [42], [43], [20], [44], [45], [46], [47], [72], [73], [49], [74], [75]。这方面的研究非常多，但是最近的神经网络设计整体倾向于越来越复杂。我们通过简单地将多尺度深度学习分为四类来继续我们的讨论：超特征学习（*Hyper Feature Learning*）、U-Net 类型、HED 类型格和 U-Ne+HED 类型。图 1中展示了它们的整体示意图。

超特征学习： 如图 1(a) 所示，超特征学习 [27], [76], [77] 是学习多尺度信息的最直观的一种方式。已有很多研究将超特征学习用于显著性物体检测 [72], [41], [69], [46], [70], [78], [71], [79]。这些模型将来自于骨干网络的不同层 [72], [41] 或多流网络的不同分支 [69], [46], [70] 的多尺度深度特征进行拼接/相加，然后将所融合的超特征（也被成为 Hypercolumn）用于最终的显著性物体预测。

U-Net 类型： 众所周知，深度神经网络的顶层包含高层语义信息，而低层则学习底层的细节特征。因此，如图 1(b) 所示，对超特征学习的一种合理改进就是将深度特征从高层到低层逐步融合 [48], [28]。以这种方式，高层的语义特征将会与低层的底层特征相结合来捕获细粒度的细节。特征融合可以是简单的逐元素求和 [48]、特征图拼接（U-Net）[28] 或基于它们的更复杂的设计。目前，许多显著性物体检测模型都是基于这种类型 [80], [81], [73], [82], [44], [43], [17], [83], [84]。这里需要注意的是，超特征学习和 U-Net 类型并没有使用深监督，因此它们是没有侧输出的。

HED 类型： HED 类型的网络 [29], [85], [86] 首先提出来是用于边缘检测的。在此之后，类似的思想也被用来进行显著性物体检测 [18], [47]。HED 类型的网络在中间层增加了深监督来获得侧输出预测，最终结果是所有侧输出预测的线性组合（如图 1(c) 所示）。与多尺度

特征融合不同，HED 使用了多尺度预测融合。

U-Net+HED 类型： U-Net+HED 类型的方法结合了 U-Net 和 HED 的优点。如图 1(d) 所示，U-Net+HED 架构是在 U-Net 解码器的每个卷积阶段都进行深监督。最近所提出的很多显著性物体检测模型都属于这一类型 [19], [42], [20], [21], [45], [87], [49], [22], [50], [51], [88], [89], [90], [91], [26]，而它们之间的不同之处在于使用不同的融合策略。这些模型的一个明显的相似之处是，最终的预测都是将侧输出预测进行线性融合而产生的。此处，多尺度学习通过以下两个方面来实现：1) U-Net 以编码-解码的形式融合从高层到低层的多层卷积特征；2) 多尺度侧输出预测被线性融合后以用于最终预测。当前，该领域的研究主要集中在第一个方面，一些性能较高的模型已经为此设计了非常复杂的特征融合策略 [20], [45]。

显著性物体检测的完整文献综述超出了本文的范围，更为详细的综述请参考 [92], [93], [94]。在本文中，我们致力于上述 U-Net+HED 多尺度学习的第二方面，即多尺度侧输出融合。我们发现，传统的线性侧输出预测融合的上限仅限于侧输出预测。因此，我们提出了一种以非线性方式融合侧输出特征的方法 DNA，从而使所融合的混合特征可以更好地利用互补的多尺度深度特征。图 1(e) 展示了所提出的 DNA 的简化图。我们证明了将 DNA 结合于非常简单的 U-Net 就可以实现较好的性能。

III. 回顾线性侧输出融合

深监督和相应的线性侧输出预测融合在许多计算机视觉任务中都被证实非常有效 [29], [85], [20], [45]。本节从理论和实验两个角度分析了线性侧输出融合的局限性。据我们所知，这是一个新的贡献。

假设一个具有深监督的网络有 N 个侧输出预测图 $\{\mathcal{O}_1, \mathcal{O}_2, \cdots, \mathcal{O}_N\}$，它们均使用真值图进行监督（如图 1(c)(d) 所示）。在不失一般性的前提下，我们假设线性侧输出融合是一个逐像素卷积，即 1×1 卷积。因此，当前的线性侧输出融合方式可以写成

$$\widehat{\mathcal{O}} = \sum_{i=1}^{N} \boldsymbol{w}_i \cdot \mathcal{O}_i, \tag{1}$$

这里，逐像素卷积的权重 $\boldsymbol{w}_i$ 是可以学习的。注意，我们有 $\boldsymbol{w}_i \geq 0$；否则，由于 $\mathcal{O}_i$ 对 $\widehat{\mathcal{O}}$ 有负面影响，因此在融合时需要将其排除在外。要获得输出的显著性概率图，还必须在 $\widehat{\mathcal{O}}$ 上使用标准的 sigmoid 函数 $\sigma(x) = \frac{1}{1+e^{-x}}$。这样，融合的概率图就变为

$$\widehat{\mathcal{P}} = \sigma(\widehat{\mathcal{O}}) = \sigma(\sum_{i=1}^{N} \boldsymbol{w}_i \cdot \mathcal{O}_i). \tag{2}$$

同样，我们可以计算侧输出概率图 $\{\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_N\}$。

理论 1. *若 $\|\boldsymbol{w}\|_1 = 1$，融合输出 $\widehat{\mathcal{P}}$ 的平均绝对误差 (Mean Absolute Error, MAE) 受侧输出预测的限制。*

证明. 若 $\|\boldsymbol{w}\|_1 = 1$，由于 $\boldsymbol{w}_i \geq 0$，所以

$$\min(\mathcal{O}_i) \leq \sum_{i=1}^{N} \boldsymbol{w}_i \cdot \mathcal{O}_i \leq \max(\mathcal{O}_i). \tag{3}$$

由于 sigmoid 函数 $\sigma(x)$ 单调递增，所以我们有

$$\min(\mathcal{P}_i) \leq \widehat{\mathcal{P}} \leq \max(\mathcal{P}_i). \tag{4}$$

若某一像素 $\boldsymbol{p}$ 为正，$\text{MAE}(\widehat{\mathcal{P}})_{\boldsymbol{p}} = |1 - \widehat{\mathcal{P}}(\boldsymbol{p})| = 1 - \widehat{\mathcal{P}}(\boldsymbol{p})$ 且 $1 - \max(\mathcal{P}_i)_{\boldsymbol{p}} \leq 1 - \widehat{\mathcal{P}}(\boldsymbol{p}) \leq 1 - \min(\mathcal{P}_i)_{\boldsymbol{p}}$，所以 $\min(\text{MAE}(\mathcal{P}_i)_{\boldsymbol{p}}) \leq \text{MAE}(\widehat{\mathcal{P}})_{\boldsymbol{p}} \leq \max(\text{MAE}(\mathcal{P}_i)_{\boldsymbol{p}})$。若某一像素 $\boldsymbol{p}$ 为负，$\text{MAE}(\widehat{\mathcal{P}})_{\boldsymbol{p}} = |0 - \widehat{\mathcal{P}}(\boldsymbol{p})| = \widehat{\mathcal{P}}(\boldsymbol{p})$ 且 $\min(\mathcal{P}_i)_{\boldsymbol{p}} \leq \widehat{\mathcal{P}}(\boldsymbol{p}) \leq \max(\mathcal{P}_i)_{\boldsymbol{p}}$, 所以 $\min(\text{MAE}(\mathcal{P}_i)_{\boldsymbol{p}}) \leq \text{MAE}(\widehat{\mathcal{P}})_{\boldsymbol{p}} \leq \max(\text{MAE}(\mathcal{P}_i)_{\boldsymbol{p}})$。需要注意的是，因为 $\boldsymbol{w}$ 通常有 N 维 (在 VGG16 [95] 和 ResNet [96] 中 $N \leq 6$)，所以很难保证上述左等式成立。因此，传统的线性融合在 MAE 度量上受到限制。然而，我们期望的却是通过充分利用多尺度信息来突破该限制。 □

引理 1. *若 $\|\boldsymbol{w}\|_1 \neq 1$，传统的线性融合（如公式 (1) 和公式 (2) 所示）等效于先使用 $\|\tilde{\boldsymbol{w}}\|_1 = 1$ 进行融合，然后使用一个单调递增的映射。*

证明. 若 $\|\boldsymbol{w}\|_1 \neq 1$，我们设置 $\boldsymbol{w} = \tilde{\boldsymbol{w}} \cdot \|\boldsymbol{w}\|_1$，因而我们有 $\|\tilde{\boldsymbol{w}}\|_1 = 1$。$\widehat{\mathcal{P}}$ 的计算变为

$$\widehat{\mathcal{P}} = \sigma(\|\boldsymbol{w}\|_1 \cdot \sum_{i=1}^{N} \tilde{\boldsymbol{w}}_i \cdot \mathcal{O}_i), \tag{5}$$

这里，$\sigma(\|\boldsymbol{w}\|_1 \cdot x)$ $(\|\boldsymbol{w}\|_1 > 0)$ 是关于 x 的单调递增函数。 □

理论 2. *$\sigma(\|\boldsymbol{w}\|_1 \cdot x)$ $(\|\boldsymbol{w}\|_1 > 0)$ 的单调递增映射无法改变 ROC 曲线和 AUC 指标[1]。*

证明. 假设正样本的预测值服从 $X \sim F(x)$ 的分布，而负样本的预测值服从 $Y \sim G(x)$ 的分布。我们可以假设

[1]AUC 是 ROC 曲线下的面积。

表 I
线性侧输出预测融合（即 LIN）和非线性侧输出特征融合（即 NONLIN）之间的比较。数据集和评测指标将在第V-A节中进行介绍。HED[29] 和 DSS[47] 的线性融合即为原论文中使用的融合方式，而它们的非线性融合就是将线性融合替换为提出的 DNA。

Methods	Fusion	DUTS-TE		ECSSD		HKU-IS		DUT-O		THUR15K	
		F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
HED [29]	linear	0.796	0.079	0.892	0.065	0.893	0.052	0.726	0.100	0.757	0.099
	nonlinear	**0.827**	**0.057**	**0.911**	**0.053**	**0.912**	**0.039**	**0.752**	**0.078**	**0.775**	**0.083**
DSS [47]	linear	0.827	0.056	0.915	0.056	0.913	0.041	0.774	0.066	0.770	0.074
	nonlinear	**0.833**	**0.055**	**0.918**	**0.056**	**0.916**	**0.040**	**0.784**	**0.060**	**0.773**	**0.072**
DNA	linear	0.844	0.048	0.921	0.050	0.917	0.034	0.765	0.066	0.785	0.071
	nonlinear	**0.865**	**0.044**	**0.935**	**0.041**	**0.930**	**0.031**	**0.799**	**0.056**	**0.793**	**0.069**

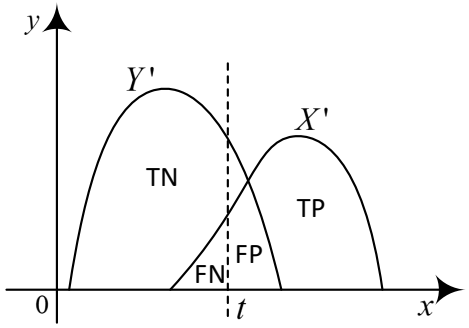

图 2. 概率（x 轴）和 X' 及 Y' 的密度（y 轴）。TN：真阴性；FN：假阴性；TP：真阳性；FP：假阳性。

F 和 G 是连续函数。$\varphi(x)=\sigma(k\cdot x)$ $(k>0)$ 是 sigmoid 函数的一种变体，因此我们有 $\varphi:\mathbb{R}\to(0,1)$ 且 φ 是单调递增函数。令 $X'=\varphi(X)$ 和 $Y'=\varphi(Y)$ 是两个变换分布，很容易得到

$$\begin{aligned}\mathbb{P}(X'\le u)&=\mathbb{P}(\varphi(X)\le u)=\mathbb{P}(X\le\varphi^{-1}(u))\\&=F(\varphi^{-1}(u)),\end{aligned}\tag{6}$$

因此我们可以得到 $X'\sim F(\varphi^{-1}(x))$ 且 $Y'\sim G(\varphi^{-1}(x))$。

令 t 为阈值，如图 2所示，真阳性率（True Positive Rate, TPR）和假阳性率（False Positive Rate, FPR）可以计算为

$$\begin{aligned}\mathrm{TPR}&=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}=\mathbb{P}(X'>t)=1-F(\varphi^{-1}(t)),\\\mathrm{FPR}&=\frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}=\mathbb{P}(Y'>t)=1-G(\varphi^{-1}(t)).\end{aligned}\tag{7}$$

因此，我们可以将 ROC 曲线表示为 $\{(1-F(\varphi^{-1}(t)),1-G(\varphi^{-1}(t))):t\in(0,1)\}$。很容易看出，随着 t 从 0 连续变化到 1，$(1-F(\varphi^{-1}(t)),1-G(\varphi^{-1}(t)))$ 也将从 $(1,1)$ 连续单调变化到 $(0,0)$。显而易见，$\{(1-F(\varphi^{-1}(t)),1-G(\varphi^{-1}(t)))\}$ 和 $\{(F(\varphi^{-1}(t)),G(\varphi^{-1}(t)))\}$ 是关于点 $(\frac{1}{2},\frac{1}{2})$ 对称的。假设曲线 $\{(1-F(\varphi^{-1}(t)),1-G(\varphi^{-1}(t)))\}$ 下的面积是 S_1 以及曲线 $\{(F(\varphi^{-1}(t)),G(\varphi^{-1}(t)))\}$ 下的面积是 S_2。通过对称，我们有 $S_1+S_2=1$。

根据以上结论，我们可以用

$$\begin{aligned}S_2&=\int_0^1 G(\varphi^{-1}(t))dF(\varphi^{-1}(t))\\&=\int_{-\infty}^{+\infty}G(x)dF(x)\end{aligned}\tag{8}$$

计算出 S_2。因此，S_2 与函数 $\varphi(x)$ 的具体形式无关，同样的，$S_1=1-S_2$ 也与 $\varphi(x)$ 的具体形式无关。此外，由于 t 的范围是 $(0,1)$，因此 $\varphi^{-1}(t)$ 的范围是 $\mathbb{R}$。我们有

$$\begin{aligned}&\{(1-F(\varphi^{-1}(t)),1-G(\varphi^{-1}(t))):t\in(0,1)\}\\&=\{(1-F(x),1-G(x)):x\in\mathbb{R}\},\end{aligned}\tag{9}$$

其同样与函数 $\varphi(x)$ 的具体形式无关。当 $F(x)$ 和 $G(x)$ 是离散的，集合 $\{(1-F(\varphi^{-1}(t)),1-G(\varphi^{-1}(t))):t\in(0,1)\}$ 也是离散的，但仍与 $\varphi(x)$ 无关。因此，我们可以得出结论，$\varphi(x)$ 无法更改 ROC 曲线和 AUC 度量。□

类似于定理 1的证明，我们可以简单地证明引理 1 的第一步，即使用 $\|\tilde{\boldsymbol{w}}\|_1=1$ 进行线性融合会限制 MAE 结果。从定理 2，我们可以得出引理 1的第二步，即一个单调递增的映射，无法改变 ROC 曲线和 AUC 值。因此，我们可以得知，传统的使用 $\|\boldsymbol{w}\|_1\neq1$ 进行线性融合的提升有限。结合定理 1，我们可以得出结论，对侧输出进行线性融合的提升效果是有限的。

除了理论上的证明，我们还进行了实验来比较显著性物体检测中的线性融合与非线性融合。为此，我们使用所提出的非线性侧输出特征融合（在第IV-B节中）进行非线性回归，来评测两个著名的模型，即 HED [29] 和 DSS [47]，以及所提出的 DNA 模型。评测结果如表 I中所示，我们可以看到从线性回归到非线性回归结果

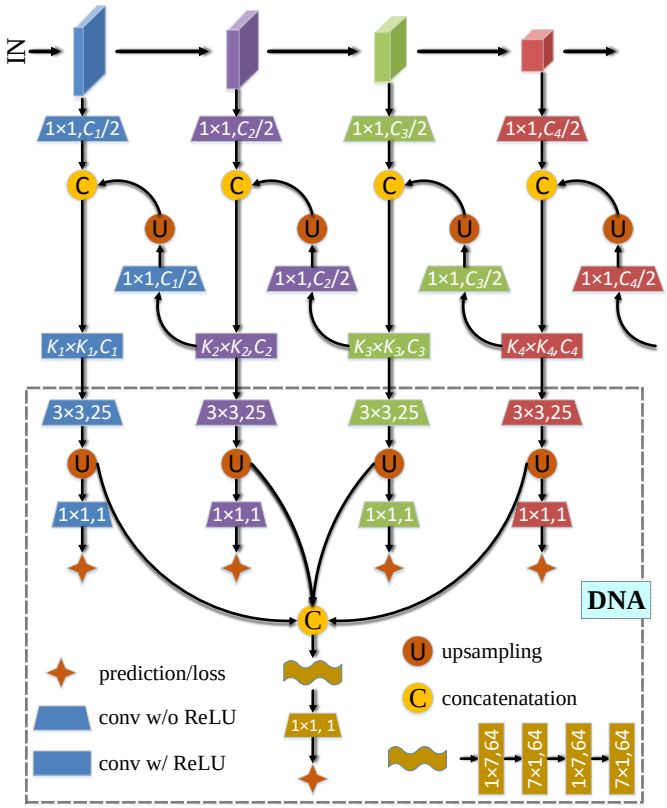

图 3. 网络架构示意图。在图中，我们仅画出前四个网络阶段，而其他两个可以用相同的方式构造。虚线框即为所提出的 DNA 模块。参数 $K_i \times K_i$ 和 C_i 将在正文中进行介绍。

有着显著提升。基于此，本文旨在设计一个简单的具有非线性侧输出融合的网络，来实现有效的显著性物体检测。

IV. 方法

本节将详细说明所提出的用于显著性物体检测的框架。首先，在第IV-A节中介绍了我们的基础网络。其次，在第IV-B节中介绍所提出的深监督非线性融合。整个网络体系结构如图 3所示。

A. 基础网络

骨干网络. 与之前的研究类似 [69], [46], [47]，我们也使用全卷积网络进行显著性物体检测。具体来说，我们使用 VGG16 网络 [95] 作为骨干网络，且为了实现“图像-图像”的转换，我们去掉了其最后的全连接层。物体检测通常需要全局信息来定位显著性物体的大致位置 [2]，所以扩大网络的感受野将有所帮助。为此，正如论文 [47] 一样，我们保留 VGG16 最后的池化层，并添加两个卷积层来代替最后的两个全连接层。其中，第一个卷积层的通道数是 $C_6^{(1)} = 192$，卷积核大小为 3×3，另一个卷积层通道数为 $C_6^{(2)} = 128$，卷积核大小为 7×7。这里，由于较大的卷积核（即 7×7）会产生更多的参数，因此我们先用 3×3 的卷积层来减少特征通道。

表 II
网络设置。

Side	C_i	$K_i \times K_i$	Resolution
Side-output 1	64	3×3	1
Side-output 2	128	3×3	1/2
Side-output 3	128	5×5	1/4
Side-output 4	128	5×5	1/8
Side-output 5	128	5×5	1/16
Side output 6	-	-	1/32

骨干网络中有五个池化层，它们将卷积层分为六个卷积块，从下到上分别表示为 $\{\mathcal{S}^1, \mathcal{S}^2, \mathcal{S}^3, \mathcal{S}^4, \mathcal{S}^5, \mathcal{S}^6\}$。我们将 $\mathcal{S}^6$ 作为高层的阀门来控制在网络中所传递的上下文信息。每个卷积块中的特征图的分辨率是前一个卷积块的一半。与之前研究相同 [47], [29]，每个卷积块的侧输出是从这个卷积块的最后一层连接出来的。

编码-解码网络. 如图 3所示，在骨干网络的基础上，我们设计了一个编码-解码网络。具体来说，我们首先在每一个卷积块 $\mathcal{S}^6$ 和 $\mathcal{S}^5$ 后连接一个 1×1 卷积层来调整通道数（如表 II所示）。然后，我们将从 $\mathcal{S}^6$ 得到的特征图上采样 2 倍。将上采样后的特征图和来自 $\mathcal{S}^5$ 的特征图进行拼接。为了融合拼接后的特征图，我们使用两个连续的卷积层来生成解码器侧端 $\widetilde{\mathcal{S}}^5$。解码器其他侧端 $\{\widetilde{\mathcal{S}}^4, \widetilde{\mathcal{S}}^3, \widetilde{\mathcal{S}}^2, \widetilde{\mathcal{S}}^1\}$ 均可以使用相同的方式获得。为清楚起见，我们上述过程表示为：

$$\begin{aligned} \widetilde{\mathcal{S}}^i &= \varphi(\mathrm{Concat}(\phi_1(\mathcal{S}^i), \phi_2(\widetilde{\mathcal{S}}^{i+1}))), \\ \phi_1(\cdot) &= \mathrm{Conv}(\cdot), \\ \phi_2(\cdot) &= \mathrm{Upsample}(\mathrm{Conv}(\cdot)), \\ \varphi(\cdot) &= \mathrm{ReLU}(\mathrm{Conv}(\cdot)), \\ \forall i &\in \{1, 2, 3, 4, 5\} \end{aligned} \tag{10}$$

注意，由于 $\mathcal{S}^6$ 是编码路径中的最后一个块，也是解码路径中的第一个块，因此我们有 $\widetilde{\mathcal{S}}^6 = \mathcal{S}^6$。通过这种方式，所提出的编码-解码网络可以把高层上下文信息传递给低层，因此较低层可以强调突出图像中显著性物体的细节。这里，解码器侧端 $\widetilde{\mathcal{S}}^i$ 的两个连续的卷积层 $(\varphi(\cdot))$ 的卷积核大小均为 $K_i \times K_i$，输出通道为 C_i。我们将在实验部分详细讨论这些参数设置。

B. 深监督非线性融合

与以前的研究 [19], [42], [20], [21], [45], [49], [22] 对多个侧输出预测使用线性融合不同，我们提出以非线性的方式融合侧输出特征。所提出的 DNA 模块如图 3的虚线框内所示。具体地说，首先，我们使用一个 3×3 卷积来为每个 $\widetilde{\mathcal{S}}^i$ 调整通道数量。然后，将特征图上采样到与原图像相同大小，以此来生成侧输出特征，而侧输出特征可以使用一个简单的 1×1 卷积来预测显著性图。在训练阶段，我们对这些预测图进行深监督。

我们将所有侧输出特征都拼接起来来构成包含丰富的多尺度和多层次信息的混合特征。非线性融合的关键思想之一是使用非对称卷积将标准二维卷积分解为两个一维卷积，即将一个 $n \times n$ 卷积分解为两个连续卷积，其卷积核大小分别为 $1 \times n$ 和 $n \times 1$。这里，我们使用不对称卷积有两个原因。一方面，在实验中，我们发现由于混合特征图具有较大的分辨率（即和原图相同的分辨率），因此在 DNA 模块中使用大卷积核可以提高性能。另一方面，对于分辨率较大的特征图而言，大卷积核的卷积是非常耗时的。根据以上分析，我们将非对称卷积的卷积核设置为 $n = 7$，而不是小的卷积核尺寸。而将卷积核设置的更大后，虽然会使准确度得到很微量的提升，但却导致计算负荷增加很多。第V-C节中我们尝试使用了不同的 n 值以及不对称/标准卷积，而结果验证了所选择的参数设置的有效性。我们用两组非对称卷积，每组由一个 1×7 和一个 7×1 卷积组成。当输入图像为 300×300 时，这些非对称卷积的 FLOP 数量（Multiply-Adds）为 13.8G，而如果使用标准的二维 7×7 卷积，则 FLOP 数量为 60.4G。最后，我们在非对称卷积后连接一个 1×1 卷积来预测最终的输出的显著性图。

训练时，我们使用类别平衡的交叉熵损失函数 [29] 来监督侧输出预测和最终的融合预测。由于 DNA 模块中的卷积层之后均连接非线性激活函数（即 ReLU），因此多尺度侧输出特征的融合是非线性的。尽管非线性函数的选择性有很多，例如 ReLU，PReLU 和 LeakyReLU 等，但在本文中，我们仅使用最常见的 ReLU 函数来证明非线性侧输出融合的有效性。传统的线性侧输出预测融合只能线性地组合多尺度预测，而所提出的非线性侧输出特征融合可以利用互补的多尺度特征来进行最终预测，因此也更加有效。相对于以前的方法，当使用第IV-A节中所描述的简单编码-解码网络时，DNA 即可实现比以前的方法更好的检测效果。值得注意的是，以前的方法 [19], [42], [20], [21] 通常设计各种网络体系结构、模块和操作来提高性能，但是在本文中，我们所提出的 DNA 仅将经过简单修改的 U-Net 作为基础网络。

V. 实验

A. 实验设置

实施细节. 关于 K_i 和 C_i 的详细设置详见表 II。由于高层使用较大的卷积核有助于提高准确性，因此，当 $i = 1, 2$ 时，$K_i \times K_i$ 等于 3×3；当 $i = 3, 4, 5$ 时，$K_i \times K_i$ 等于 5×5。对于 $i = 1, \cdots, 5$，C_i 的值分别为 64、128、128、128 和 128。在测试阶段，由于没有使用侧输出预测结果，因此我们就删除这些侧输出预测模块。而在训练阶段，深监督可以帮助训练并提高最终显著性预测的准确性，因此我们将其保留（将在第V-C节中进行证实）。

我们使用 Caffe [97] 框架来实现所提出的网络。最初的 VGG16 [95] 中包含的卷积层使用公开的经过 ImageNet 预训练的模型 [98] 进行初始化。其他层的权重初始化服从标准差为 0.01 的零均值高斯分布，偏差初始化为 0。上采样操作由具有双线性插值核的反卷积层实现，该双线性插值内核在训练过程中冻结。由于反卷积层不需要训练，因此在计算参数数量时可以直接将其忽略。整个网络使用 SGD 进行优化，学习率策略为 *poly*，即当前学习率等于初始学习率乘以 $(1 - curr\_iter/max\_iter)^{power}$。其中，超参数 *power* 和 *max_iter* 分别设置为 0.9 和 20000，因此总共需要训练 20000 次迭代。初始学习率设置为 1e-7，这也是防止网络在训练时梯度爆炸的最大值（较大的初始学习率会导致 “Nan” 错误）。我们遵循之前的显著性物体检测方法 [47], [19], [20], [18], [49], [67], [44], [50], [59], [60], [65], [68], [70], [69], [72], [78], [90] 来将动量和权重衰减分别设为经典的 0.9 和 0.0005 [99], [95]。本文的所有实验均在一块 TITAN Xp GPU 上实现。

数据集. 我们在六个最常用的数据集上评测了我们的方法，包括 DUTS [100]、ECSSD [101]、SOD [102]、HKU-IS [66]、THUR15K [103] 和 DUT-O (即 DUT-OMRON) [57]。这六个数据集分别由 15572、1000、300、4447、6232 和 5168 张复杂的自然图像组成，并均带有相应的标记好的像素级真值图。其中，DUTS 数据集 [100] 是由在非常复杂的场景中的 10553 张训练图像和 5019 张测试图像组成。为了公平比较，我们也像之前的研究一

表 III

所提出的 DNA 与 16 个显著性物体检测模型在六个数据集上关于 F_β-measure 和 MAE 两个指标的检测结果比较。我们展示分别以 VGG16 [95] 和以 ResNet-50 [96] 作为骨干网络的检测结果。每一列中，检测结果最好的前三个模型分别用 **红色**、**绿色**和**蓝色**突出显示。而对于基于 ResNet-50 的方法，我们仅突出显示达到最好性能的模型。

Methods	DUTS-TE		ECSSD		HKU-IS		DUT-O		SOD		THUR15K	
	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
Non-deep learning												
DRFI [30]	0.649	0.154	0.777	0.161	0.774	0.146	0.652	0.138	0.704	0.217	0.670	0.150
VGG16 [95] backbone												
MDF [66]	0.707	0.114	0.807	0.138	-	-	0.680	0.115	0.764	0.182	0.669	0.128
LEGS [65]	0.652	0.137	0.830	0.118	0.766	0.119	0.668	0.134	0.733	0.194	0.663	0.126
DCL [72]	0.785	0.082	0.895	0.080	0.892	0.063	0.733	0.095	0.831	0.131	0.747	0.096
DHS [49]	0.807	0.066	0.903	0.062	0.889	0.053	-	-	0.822	0.128	0.752	0.082
ELD [67]	0.727	0.092	0.866	0.081	0.837	0.074	0.700	0.092	0.758	0.154	0.726	0.095
RFCN [80]	0.782	0.089	0.896	0.097	0.892	0.080	0.738	0.095	0.802	0.161	0.754	0.100
NLDF [73]	0.806	0.065	0.902	0.066	0.902	0.048	0.753	0.080	0.837	0.123	0.762	0.080
DSS [47]	0.827	**0.056**	0.915	**0.056**	**0.913**	**0.041**	**0.774**	**0.066**	**0.842**	**0.122**	0.770	**0.074**
Amulet [45]	0.778	0.085	0.913	0.061	0.897	0.051	0.743	0.098	0.795	0.144	0.755	0.094
UCF [81]	0.772	0.112	0.901	0.071	0.888	0.062	0.730	0.120	0.805	0.148	0.758	0.112
PiCA [20]	**0.837**	**0.054**	**0.923**	**0.049**	**0.916**	**0.042**	0.766	0.068	0.836	**0.102**	**0.783**	0.083
C2S [17]	0.811	0.062	0.907	0.057	0.898	0.046	0.759	0.072	0.819	**0.122**	**0.775**	0.083
RAS [18]	**0.831**	0.059	**0.916**	0.058	**0.913**	0.045	**0.785**	**0.063**	**0.847**	0.123	0.772	**0.075**
DNA	**0.865**	**0.044**	**0.935**	**0.041**	**0.930**	**0.031**	**0.799**	**0.056**	**0.853**	**0.107**	**0.793**	**0.069**
ResNet-50 [96] backbone												
SRM [46]	0.826	0.059	0.914	0.056	0.906	0.046	0.769	0.069	0.840	0.126	0.778	0.077
BRN [44]	0.827	0.050	0.919	0.043	0.910	0.036	0.774	0.062	0.843	**0.103**	0.769	0.076
PiCA [20]	0.853	0.050	0.929	0.049	0.917	0.043	0.789	0.065	0.852	**0.103**	0.788	0.081
DNA	**0.873**	**0.040**	**0.938**	**0.040**	**0.934**	**0.029**	**0.805**	**0.056**	**0.855**	0.110	**0.796**	**0.068**

样 [44], [20], [46], [41]，将 DUTS 训练集用于模型训练，将 DUTS 测试集（DUTS-TE）和其他数据集用于测试。

评测指标. 我们使用三个评测指标来评测我们的方法以及其他最近被提出的显著性物体检测器。这三个评测指标分别为最大 F_β 值（F_β-measure）、平均绝对误差（Mean Absolute Error, MAE）和加权 F_β^ω 值（F_β^ω-measure）[104]。

给定预测出的具有连续概率值的显著性图，我们可以通过选择一个阈值将其转换为二值图并计算相应的准确率/召回率（Precision/Recall）。通过取整个数据集所有图像的准确率/召回率的平均值，我们可以获得许多平均准确率/召回率对。而 F_β-measure 是一个总性能指标：

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \tag{11}$$

其中，β^2 通常设置为 0.3 来强调精度。根据之前的研究 [73], [47], [45], [81], [20], [17], [18]，我们计算不同阈值下的 F_β-measure 的最大值。

将给定的显著性图 S 和相应的真值图 G 归一化为 [0，1] 后，MAE 可计算为

$$\text{MAE} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} |S(i,j) - G(i,j)|, \tag{12}$$

其中，H 和 W 分别表示高度和宽度，$S(i,j)$ 表示在位置 (i,j) 的显著性值，$G(i,j)$ 的定义与 $S(i,j)$ 相同。

正如 [104] 中所表明的，传统的评测指标很容易受到插值缺陷、依赖缺陷和等价缺陷的影响。基于此，F_β^ω-measure 被提出来以弥补这些缺陷。我们遵循 [52], [82], [53], [20] 来使用 F_β^ω-measure 作为指标，并使用其默认设置 [104]。

B. 性能比较

我们将所提出的显著性物体检测器与最近的 16 种具有竞争力的显著性物体检测模型进行了比较，这 16 种模型分别为 DRFI [30]、MDF [66]、LEGS [65]、DCL [72]、DHS [49]、ELD [67]、RFCN [80]、NLDF [73]、DSS [47]、SRM [46]、Amulet [45]、UCF [81]、BRN [44]、PiCA [20]、C2S [17] 和 RAS [18]。其中，DRFI[30] 是最著名的基于非深度学习的方法，而其他 15 种均是基于深

表 IV
所提出的 DNA 与 16 个显著性物体检测模型在六个数据集上关于 F_β^ω-MEASURE 指标的检测结果比较。参数数量（#PARAM）的单位为百万（M），速度的单位为帧每秒（FPS）。我们展示分别以 VGG16 [95] 和以 RESNET-50 [96] 作为骨干网络的检测结果。每一列中，检测结果最好的前三个模型分别用 **红色**、**绿色**和**蓝色**突出显示。而对于基于 RESNET-50 的方法，我们仅突出显示达到最好性能的模型。

Methods	#Param	Speed	DUTS-TE	ECSSD	HKU-IS	DUT-O	SOD	THUR15K
Non-deep learning								
DRFI [30]	-	1/8	0.378	0.548	0.504	0.424	0.450	0.444
VGG16 [95] backbone								
MDF [66]	56.86	1/19	0.507	0.619	-	0.494	0.528	0.508
LEGS [65]	**18.40**	0.6	0.510	0.692	0.616	0.523	0.550	0.538
DCL [72]	66.24	1.4	0.632	0.782	0.770	0.584	0.669	0.624
DHS [49]	94.04	10.0	0.705	0.837	0.816	-	0.685	0.666
ELD [67]	43.09	1.0	0.607	0.783	0.743	0.593	0.634	0.621
RFCN [80]	134.69	0.4	0.586	0.725	0.707	0.562	0.591	0.592
NLDF [73]	35.49	**18.5**	0.710	0.835	0.838	0.634	0.708	0.676
DSS [47]	62.23	7.0	0.700	0.832	0.821	0.643	0.698	0.662
Amulet [45]	33.15	9.7	0.657	0.839	0.817	0.626	0.674	0.650
UCF [81]	23.98	12.0	0.595	0.805	0.779	0.574	0.673	0.613
PiCA [20]	32.85	5.6	**0.745**	**0.862**	**0.847**	**0.691**	**0.721**	**0.688**
C2S [17]	137.03	16.7	0.717	0.849	0.835	0.663	0.700	0.685
RAS [18]	**20.13**	**20.4**	**0.739**	**0.855**	**0.850**	**0.695**	**0.718**	**0.691**
DNA	**20.06**	**25.0**	**0.797**	**0.897**	**0.889**	**0.729**	**0.755**	**0.723**
ResNet-50 [96] backbone								
SRM [46]	43.74	12.3	0.721	0.849	0.835	0.658	0.670	0.684
BRN [44]	126.35	3.6	0.774	0.887	0.876	0.709	0.738	0.712
PiCA [20]	37.02	4.4	0.754	0.863	0.841	0.695	0.722	0.690
DNA	**29.31**	**12.8**	**0.810**	**0.901**	**0.898**	**0.735**	**0.755**	**0.730**

度学习的模型。由于 MDF [66] 使用了 HKU-IS 数据集 [66] 中的部分数据进行训练，我们没有报告 MDF 在 HKU-IS 数据集上的结果。同样地，我们没有报告 DHS [49] 在 DUT-O 数据集 [57] 上的结果。由于 SRM [46] 和 BRN [44] 是基于 ResNet-50 [96] 骨干网络构建的，为了公平比较，我们还评测了基于 ResNet-50 版本 DNA 的和 PiCA [20] 模型。所有以前的方法都使用其公开代码和作者公布的预训练模型以及默认设置进行测试。

F_β-measure 和 MAE. 表 III总结了 F_β-measure 和 MAE 在六个数据集上的评测结果。在大多数情况下，DNA 的性能都优于其他检测模型，因此可以证明其有效性。使用 VGG16[95] 作为骨干网络时，在 DUTS-TE、ECSSD、HKU-IS、DUT-O、SOD 和 THUR15K 六个数据集上，DNA 的 F_β-measure 比次优方法分别高 2.8%、1.2%、1.4%、1.4%、0.6% 和 1.0%。关于 MAE 指标，除了在 SOD 数据集上，DNA 比 PiCA [20] 性能稍差以外，DNA 也都达到最好结果。总体而言，PiCA [20] 是除 DNA 以外的最优模型。当使用 ResNet-50 作为骨干网络时，DNA 仍然比之前的显著性物体检测模型性能更好。这说明 DNA 对不同的网络体系结构都非常鲁棒。因此，我们建议未来的显著性物体检测模型均使用非线性侧输出融合来代替传统的线性融合。

F_β^ω-measure. F_β^ω-measure 也是一个常用的显著性评测指标。在表 IV中，我们使用 F_β^ω-measure 评测 DNA 和其他检测模型。在 DUTS-TE、ECSSD、HKU-IS、DUT-O、SOD 和 THUR15K 六个数据集上，VGG16 版本的 DNA 在 F_β^ω-measure 上比次优模型的结果分别高 5.2%、3.5%、3.9%、3.4%、3.4% 和 3.2%。对于 ResNet-50 版本，DNA 的 F_β^ω-measure 比以前的最好模型高 3.6%、1.4%、2.2%、2.6%、1.7% 和 1.8%。此外，DNA 的网络体系结构非常简单，因此易于扩展并用于其他计算机视觉应用。

参数数量和运行时间. 如表 IV所示，DNA 具有较少的参数。具体而言，VGG16 版本 DNA 的参数约 20M，ResNet-50 版本 DNA 参数约 29M。并且，DNA 的运行速度也比其他方法快。对于 VGG16 版本，运行速度达到 25fps，ResNet-50 版本也可达到 12.8fps。

定性比较. 如图 4所示，为了从视觉上直观的表明 DNA

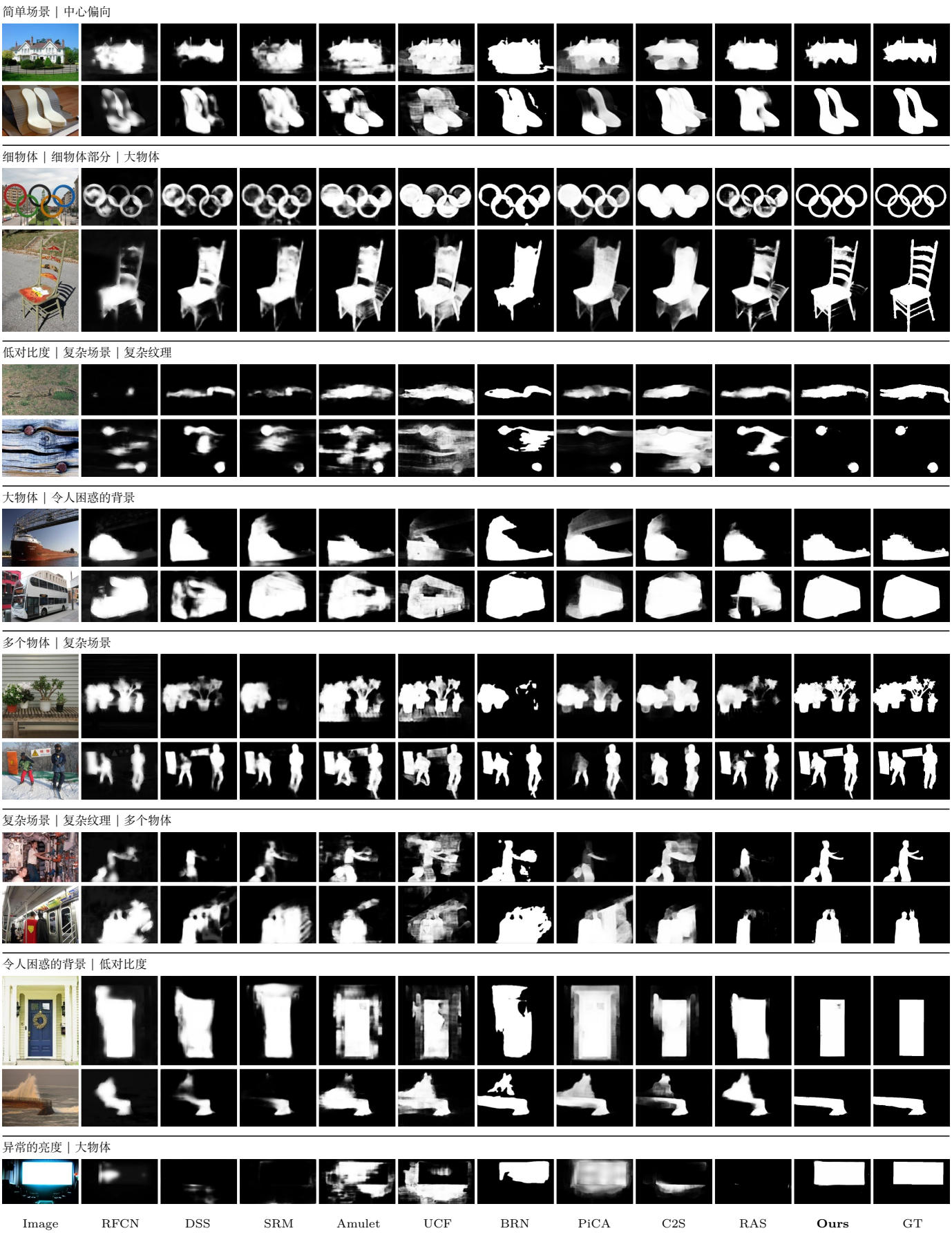

图 4. DNA 与最新显著性检测模型的定性比较结果。此处 GT 为真值图。

表 V
消融实验。U-Net 是指使用 VGG16 作为骨干网络的标准 U-Net [28]。如果删除 DNA 模块和深监督，那么所提出的网络（图 3中）就变为一个编码-解码网络，即 *Encoder-Decoder*。如果进一步将 Encoder-Decoder 高层的所有卷积替换为 3×3 卷积，就得到 *Encoder-Decoder w/ K3*。*Encoder-Decoder w/ lin* 是将图 3中的 DNA 模块替换为论文 [29] 中的传统线性融合。

Methods	DUTS-TE		ECSSD		HKU-IS		DUT-O		SOD		THUR15K	
	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
U-Net	0.793	0.080	0.890	0.065	0.894	0.051	0.723	0.101	0.811	0.115	0.758	0.099
Encoder-Decoder w/ K3	0.766	0.101	0.869	0.081	0.876	0.064	0.687	0.129	0.778	0.131	0.736	0.112
Encoder-Decoder	0.831	0.053	0.911	0.052	0.916	0.037	0.754	0.073	0.830	0.117	0.780	0.077
Encoder-Decoder w/ lin	0.844	0.048	0.921	0.050	0.917	0.034	0.765	0.066	0.839	0.120	0.785	0.071
DNA w/o Deep Supervision	0.867	0.042	0.932	0.041	0.927	0.032	0.788	0.059	0.860	0.103	0.794	0.068
DNA	0.865	0.044	0.935	0.041	0.930	0.031	0.799	0.056	0.853	0.107	0.793	0.069

表 VI
各种参数设置的消融实验。参数数量（#Param）的单位为百万（M），速度的单位为帧每秒（fps）。

Methods	#Param	Speed	DUTS-TE		ECSSD		HKU-IS		DUT-O		SOD		THUR15K	
			F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
#1	18.49	27.0	0.859	0.044	0.932	0.041	0.928	0.031	0.796	0.057	0.855	0.105	0.790	0.069
#2	20.06	25.0	0.865	0.044	0.935	0.041	0.930	0.031	0.799	0.056	0.853	0.107	0.793	0.069
#3	27.88	22.7	0.866	0.043	0.936	0.041	0.930	0.031	0.799	0.056	0.861	0.106	0.792	0.069
#4	41.41	18.2	0.864	0.044	0.935	0.041	0.931	0.030	0.800	0.056	0.857	0.105	0.792	0.069

表 VII
表 VI中参数设置的消融实验。深色突出显示的为本文的默认设置。

No.	#1	#2	#3	#4
Side 1	$(3 \times 3, 64)$	$(3 \times 3, 64)$	$(3 \times 3, 64)$	$(3 \times 3, 64)$
Side 2	$(3 \times 3, 128)$	$(3 \times 3, 128)$	$(3 \times 3, 128)$	$(3 \times 3, 128)$
Side 3	$(3 \times 3, 128)$	$(5 \times 5, 128)$	$(5 \times 5, 128)$	$(5 \times 5, 256)$
Side 4	$(3 \times 3, 128)$	$(5 \times 5, 128)$	$(5 \times 5, 256)$	$(5 \times 5, 256)$
Side 5	$(3 \times 3, 128)$	$(5 \times 5, 128)$	$(5 \times 5, 256)$	$(5 \times 5, 512)$
Side 6	$(192, 128)$	$(192, 128)$	$(256, 256)$	$(256, 256)$

相较于之前方法的优势，我们从各数据集中选择了一些具有代表性的图像进行定性比较。所选择的图像尽量包含各种不同的场景，例如复杂场景、显著性物体结构较细、前景和背景之间对比度低、包含多个不同尺寸的物体、场景亮度异常等。在图 4中，我们将所选图像分为多个组，每组带有多个标记以描述其属性。考虑到所有情况，即使在复杂、低对比度以及异常的场景中，所提出的 DNA 都能正确分割出具有连贯边界和连通区域的显著性物体。这也正是在上述定量比较中，DNA 性能优于其他方法的原因。

C. 消融实验

非线性融合和线性融合. 为了证明非线性融合的有效性，我们通过用传统的线性侧输出预测融合 [29] 替换我们网络中的 DNA 模块来获得一个新的深监督的编码-解码网络，即 *Encoder-Decoder w/lin*。结果如**表 V**所示，我们可以清楚地看到，就 F_β-measure 和 MAE 而言，非线性融合的效果都明显好于线性融合。线性侧输出融合和非线性融合的定性比较如图 5中的第 4 列和第 5 列所示。非线性融合在各种复杂场景下都表现出绝对的优越性。

所提出的编码-解码架构和标准的 U-Net. 如果移除 DNA 模块和深监督，那么所提出的编码-解码架构就变成了简单修改版本的 U-Net [28]。首先，我们将位于高层的所有卷积的卷积核大小，即 $K_3 \times K_3$、$K_4 \times K_4$ 和 $K_5 \times K_5$，全部改为 3×3。如表 V所示，得到的模型 *Encoder-Decoder w/ K3* 的性能不如标准 U-Net [28]。这可能是因为所提出的编码-解码架构具有更少的特征通道数，因此就具有更少的参数（U-Net 有 31.06M 参数）。接下来，我们在高层使用默认的卷积核大小 5×5。所得到的模型 *Encoder-Decoder* 的性能优于 U-Net。这说明在高层使用较大的卷积核对于提高性能很重要。图 5中的第 2 列和第 3 列展示了 *Encoder-Decoder* 与 U-Net 的定性比较。可以看出，*Encoder-Decoder* 可以预测出更好的显著性图。

有/无深监督的编码-解码架构. 在表 V中，*Encoder-Decoder w/ lin* 性能要优于 *Encoder-Decoder*，这从图 5中的第 3 列和第 4 列同样也可以看出来。如果移除了 DNA 里的深监督，所得到的模型 *DNA w/o Deep*

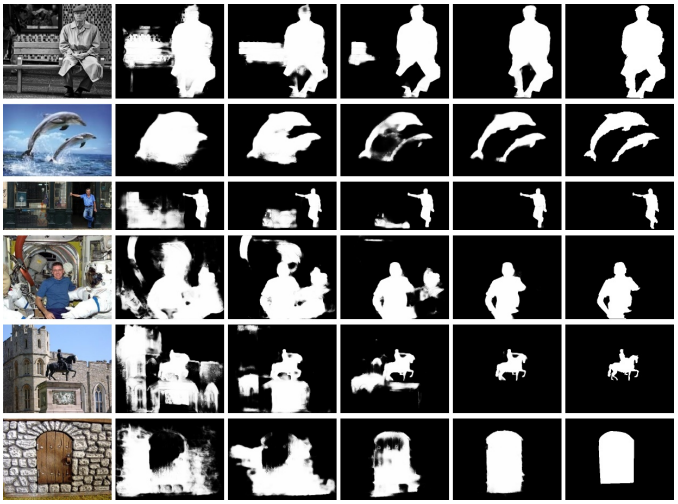

图 5. 不同模型变体之间的定性比较。ED：Encoder-Decoder，ED w/ lin：Encoder-Decoder w/ lin。从此图可以清楚地看到，显著性预测的质量从左到右逐渐提高。由于 *ED w/ lin* 仅仅将 DNA 中的非线性侧输出融合替换为线性融合，因此该图可以证明非线性融合在显著性检测中的优越性。

表 VIII
DNA 模块中的各种卷积核大小。

Datasets	Metrics	3×3	5×5	7×7	1×7 7×1	1×9 9×1
DUTS-TE	F_β	0.861	0.863	0.865	0.865	0.864
	MAE	0.045	0.045	0.043	0.044	0.044
ECSSD	F_β	0.930	0.933	0.935	0.935	0.935
	MAE	0.042	0.041	0.041	0.041	0.040
DUT-O	F_β	0.795	0.797	0.799	0.799	0.798
	MAE	0.058	0.058	0.057	0.056	0.056
Speed (fps)		27.8	23.2	19.6	25.0	20.4

Supervision 在大多数情况下都比原本的 DNA 性能差。因此，深监督可以明显地改善显著性预测性能。

参数设置. 为了评测不同参数设置的影响，我们尝试了表 VII 中各种不同的参数设置。对于第 1-5 侧端，我们展示 $(K_i \times K_i, C_i)$ 的设置。对于第 6 层，我们展示 $C_6^{(1)}, C_6^{(2)}$ 的设置。评测结果如表 VI所示。从第一、第二个实验中，我们可以看出高层使用较大的卷积核会产生更好的结果，但提升程度不如表 V中的深监督显著。从第三、第四个实验中，我们发现通过增加卷积层通道数来引入更多参数，可以产生略好的结果。考虑到性能、参数数量和速度之间的权衡，我们选择第二组设置作为默认参数设置。

DNA 模块中的不对称卷积. 在表 VIII中，我们评测 DNA 模型的各种不同的卷积核大小。总体上大卷积核的性能要好于小卷积核，但将卷积核从 7 增加到 9 却没有提高性能。如表 VIII所示，由于 DNA 中的特征图具有与原始图像相同的分辨率，标准的二维 7×7 卷积是非常耗时的。因此，我们使用非对称卷积（即 $1\times7, 7\times1$）来同时实现较大的卷积核和较快的速度。

VI. 总结

之前的深监督显著性物体检测网络使用线性侧输出预测融合，我们在理论上和实验上均证明了线性侧输出融合是次优的且不如非线性融合。基于此，我们提出了以非线性方式融合多层次的侧输出特征的 DNA 模块。与 16 个最近的显著性检测模型相比时，当用于经简单修改过的 U-Net，DNA 可以在各种指标下达到新的最高水平。所提出的网络还具有更少的参数和更快的运行速度，这更加证明了其有效性。在以后的研究中，我们计划将 DNA 用于进一步改进显著性物体检测，并将其用于需要多尺度和多层次信息的其他计算机视觉任务中。

参考文献

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1597–1604.

[2] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2015.

[3] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational Visual Media*, vol. 5, no. 2, pp. 117–150, 2019.

[4] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 363–376, 2013.

[5] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2Photo: Internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 124:1–10, 2009.

[6] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009.

[7] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, and F. Porikli, "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3516–3527, 2019.

[8] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition." *IEEE Trans. Circ. Syst. Video Technol.*, vol. 24, no. 5, pp. 769–779, 2014.

[9] F. Zund, Y. Pritch, A. Sorkine-Hornung, S. Mangold, and T. Gross, "Content-aware compression using saliency-driven image retargeting," in *IEEE Int. Conf. Image Process.*, 2013, pp. 1845–1849.

[10] M.-M. Cheng, X.-C. Liu, J. Wang, S.-P. Lu, Y.-K. Lai, and P. L. Rosin, "Structure-preserving neural style transfer," *IEEE Trans. Image Process.*, vol. 29, pp. 909–920, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8816670

[11] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Trans. Visual. Comput. Graph.*, vol. 23, no. 8, pp. 2014–2027, 2016.

[12] W. Wang, J. Shen, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, 2018.

[13] J. Shen, J. Peng, and L. Shao, "Submodular trajectories for better motion segmentation in videos," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2688–2700, 2018.

[14] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, 2018.

[15] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 1568–1576.

[16] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7268–7277.

[17] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *Eur. Conf. Comput. Vis.*, 2018, pp. 355–370.

[18] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Eur. Conf. Comput. Vis.*, 2018.

[19] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 714–722.

[20] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3089–3098.

[21] M. A. Islam, M. Kalash, and N. D. Bruce, "Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7142–7150.

[22] S. He, J. Jiao, X. Zhang, G. Han, and R. W. Lau, "Delving into salient object subitizing and detection," in *Int. Conf. Comput. Vis.*, 2017, pp. 1059–1067.

[23] S. Wang, S. Yang, M. Wang, and L. Jiao, "New contour cue-based hybrid sparse learning for salient object detection," *IEEE Trans. on Cybernetics*, 2019.

[24] K. Yan, X. Wang, J. Kim, and D. Feng, "A new aggregation of DNN sparse and dense labeling for saliency detection," *IEEE Trans. on Cybernetics*, 2020.

[25] H. Li, G. Li, B. Yang, G. Chen, L. Lin, and Y. Yu, "Depthwise nonlocal module for fast salient object detection using a single thread," *IEEE Trans. on Cybernetics*, 2020.

[26] S. Chen, B. Wang, X. Tan, and X. Hu, "Embedding attention and residual network for accurate salient object detection," *IEEE Trans. Cybernetics*, vol. 50, no. 5, pp. 2050–2062, 2020.

[27] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 447–456.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[29] S. Xie and Z. Tu, "Holistically-nested edge detection," *Int. J. Comput. Vis.*, vol. 125, no. 1-3, pp. 3–18, 2017.

[30] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," *Int. J. Comput. Vis.*, vol. 123, no. 2, pp. 251–268, 2017.

[31] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1884–1892.

[32] L. Zhu, H. Ling, J. Wu, H. Deng, and J. Liu, "Saliency pattern detection by ranking structured trees," in *Int. Conf. Comput. Vis.*, 2017, pp. 5467–5476.

[33] F. Guo, W. Wang, J. Shen, L. Shao, J. Yang, D. Tao, and Y. Y. Tang, "Video saliency detection using object proposals," *IEEE Trans. Cybernetics*, vol. 48, no. 11, pp. 3159–3170, 2017.

[34] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, 2017.

[35] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, 2017.

[36] W. Wang, J. Shen, F. Porikli, and R. Yang, "Semi-supervised video object segmentation with super-trajectories," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 985–998, 2018.

[37] R. Fan, M.-M. Cheng, Q. Hou, T.-J. Mu, J. Wang, and S.-M. Hu, "S4Net: Single stage salient-instance segmentation," *Computational Visual Media*, vol. 6, no. 2, pp. 191–204, June 2020.

[38] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, 2021.

[39] Y. Liu, S.-J. Li, and M.-M. Cheng, "RefinedBox: Refining for fewer and high-quality object proposals," *Neurocomputing*, vol. 406, pp. 106–116, 2020.

[40] M.-M. Cheng, Y. Liu, W.-Y. Lin, Z. Zhang, P. L. Rosin, and P. H. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," *Computational Visual Media*, vol. 5, no. 1, pp. 3–20, 2019.

[41] Y. Zeng, H. Lu, L. Zhang, M. Feng, and A. Borji, "Learning to promote saliency detectors," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 1644–1653.

[42] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 1711–1720.

[43] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 1741–1750.

[44] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3127–3135.

[45] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

[46] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Int. Conf. Comput. Vis.*, 2017, pp. 4019–4028.

[47] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, 2019.

[48] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3431–3440.

[49] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 678–686.

[50] Z. Li, C. Lang, Y. Chen, J. Liew, and J. Feng, "Deep reasoning with multi-scale context for salient object detection," *arXiv preprint arXiv:1901.08362*, 2019.

[51] S. Jia and N. D. Bruce, "Richer and deeper supervision network for salient object detection," *arXiv preprint arXiv:1901.02425*, 2019.

[52] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, and J. Yang, "Saliency propagation from simple to difficult," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 2531–2539.

[53] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2334–2342.

[54] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 4321–4329.

[55] J. Shen, J. Peng, X. Dong, L. Shao, and F. Porikli, "Higher order energies for image segmentation," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4911–4922, 2017.

[56] Z. Jiang and L. S. Davis, "Submodular salient region detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 2043–2050.

[57] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3166–3173.

[58] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2814–2821.

[59] D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in *Int. Conf. Comput. Vis.*, 2017, pp. 4048–4056.

[60] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley, "Deep unsupervised saliency detection: A multiple noisy labeling perspective," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 9029–9038.

[61] H. R. Tavakoli, F. Ahmed, A. Borji, and J. Laaksonen, "Saliency revisited: Analysis of mouse movements versus fixations," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 6354–6362.

[62] C. Lang, J. Feng, S. Feng, J. Wang, and S. Yan, "Dual low-rank pursuit: Learning salient features for saliency detection," *IEEE Trans. Neural Net. Learn. Syst.*, vol. 27, no. 6, pp. 1190–1200, 2016.

[63] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, 2016.

[64] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1265–1274.

[65] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 3183–3192.

[66] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 5455–5463.

[67] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 660–668.

[68] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Net. Learn. Syst.*, vol. 27, no. 6, pp. 1135–1149, 2016.

[69] X. Chen, A. Zheng, J. Li, and F. Lu, "Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic CNNs," in *Int. Conf. Comput. Vis.*, 2017, pp. 1050–1058.

[70] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3264–3274, 2018.

[71] N. Liu, J. Han, T. Liu, and X. Li, "Learning to predict eye fixations via multiresolution convolutional neural networks," *IEEE Trans. Neural Net. Learn. Syst.*, vol. 29, no. 2, pp. 392–404, 2018.

[72] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 478–487.

[73] Z. Luo, A. K. Mishra, A. Achkar, J. A. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 6609–6617.

[74] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3907–3916.

[75] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "CapSal: Leveraging captioning to boost semantics for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 6024–6033.

[76] X. Hu, Y. Liu, K. Wang, and B. Ren, "Learning hybrid convolutional features for edge detection," *Neurocomputing*, vol. 313, pp. 377–385, 2018.

[77] Y. Liu, P.-T. Jiang, V. Petrosyan, S.-J. Li, J. Bian, L. Zhang, and M.-M. Cheng, "DEL: Deep embedding learning for efficient image segmentation." in *Int. Joint Conf. Artif. Intell.*, 2018, pp. 864–870.

[78] J. Su, J. Li, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 3799–3808.

[79] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3085–3094.

[80] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Eur. Conf. Comput. Vis.*, 2016, pp. 825–841.

[81] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Int. Conf. Comput. Vis.*, 2017, pp. 212–221.

[82] P. Hu, B. Shuai, J. Liu, and G. Wang, "Deep level sets for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2300–2309.

[83] N. D. Bruce, C. Catton, and S. Janjic, "A deeper look at saliency: Feature contrast, semantics, and beyond," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 516–524.

[84] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 7479–7489.

[85] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1939–1946, 2019.

[86] Y. Liu, M.-M. Cheng, D.-P. Fan, L. Zhang, J. Bian, and D. Tao, "Semantic edge detection with diverse deep supervision," *arXiv preprint arXiv:1804.02864*, 2018.

[87] Y. Qiu, Y. Liu, S. Li, and J. Xu, "MiniSeg: An extremely minimum network for efficient COVID-19 segmentation," in *AAAI Conf. Artif. Intell.*, 2021.

[88] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 1448–1457.

[89] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 1623–1632.

[90] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3917–3926.

[91] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8150–8159.

[92] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.

[93] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circ. Syst. Video Technol.*, 2018.

[94] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 84–100, 2018.

[95] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent.*, 2015.

[96] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 770–778.

[97] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[98] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.

[99] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Adv. Neural Inform. Process. Syst.*, 2012, pp. 1097–1105.

[100] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 136–145.

[101] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 1155–1162.

[102] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2010, pp. 49–56.

[103] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu, "SalientShape: Group saliency in image collections," *The Visual Computer*, vol. 30, no. 4, pp. 443–453, 2014.

[104] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 248–255.