Deep Unsupervised Saliency Detection: A Multiple Noisy Labeling Perspective

这篇论文的角度很新颖，从Multiple Noisy Labeling的视角考虑Deep Unsupervised Saliency Detection。
当前Deep Saliency Detection方法的成功在很大程度上取决于每像素标记形式的大规模监督的可用性。这种监督通常需要大量的标注(labor-intensive)而且并非总是可能，往往会妨碍学习模型的泛化能力。这引出了一个自然的问题：“是否有可能在不使用标记数据的情况下学习saliency maps而改善泛化能力？”自然就会想到Unsupervised，基于Unsupervised Saliency Detection方法的传统手工特征，即使已经被deep supervised methods超越，虽然在性能上比不过监督方法，但是通常是数据集独立的并且可以应用到自然环境中。
一些现有的非监督方法的结果虽然有噪声，但是却包含有用的信息。为此，本文通过学习由“weak”和“noisy” unsupervised handcrafted saliency methods产生的多个噪声标记，对unsupervised saliency detection提出了一种新的视角。
本文用于unsupervised saliency detection的端到端深度学习框架包括latent saliency prediction module和noise modeling module，它们协同工作并联合优化。显式噪声建模使我们能够以概率方式处理噪声显著图。各种benchmarking数据集的广泛实验结果表明，本文的模型不仅优于所有unsupervised saliency detection方法，而且具有较大的优势，而且与最近的supervised deep saliency方法相比也达到了相当的性能。
本文的一个主要思路就是从问题的本质出发，认为尽管一些非监督的方法不准确，但是实质上应该转化为真值与noise的组合，然后分别进行建模。建模方式中通过使用cross entropy损失，然后想办法让noise进行逼近就可以了。

Framework

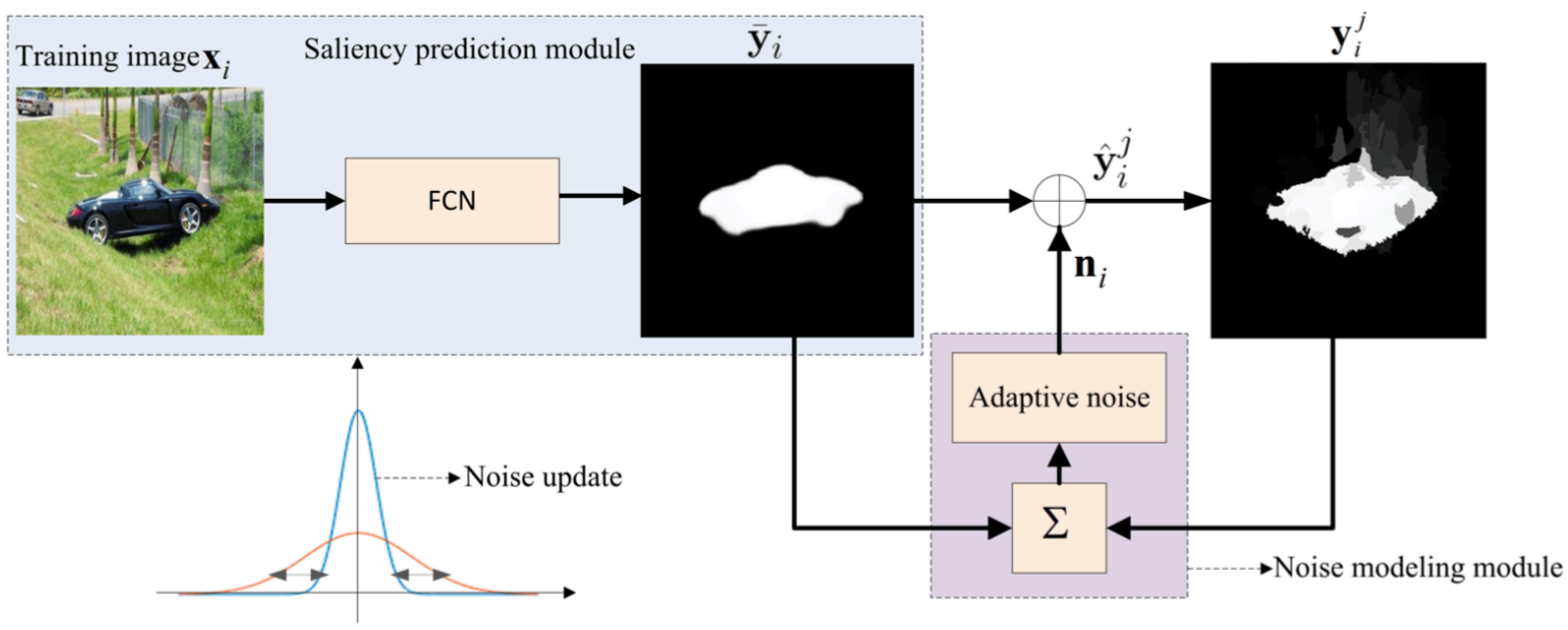

Figure 2. Conceptual illustration of our saliency detection framework, which consists of a “latent” saliency prediction module and a noise modeling module. Given an input image, noisy saliency maps are generated by handcrafted feature based unsupervised saliency detection methods. Our framework jointly optimizes both modules under a unified loss function. The saliency prediction module targets at learning latent saliency maps based on current noise estimation and the noisy saliency maps. The noise modeling module updates the noise estimation in different saliency maps based on updated saliency prediction and the noisy saliency maps. In our experiments, the overall optimization converges in several rounds.

本文提出一种新颖的unsupervised saliency detection的端到端深度学习框架，从别的非监督方法产生的结果进行学习，这些结果通常带有noise，因此本文提出y = y' + n的思想，学习一个潜在的显著性预测模块和一个噪声模块。

损失函数的设计

分成两部分，最终的损失函数，是显著性预测模块的损失函数加上噪声模块的损失函数，同时用一个正则化因子进行权衡。

$$\hat{\mathbf{y}}_i^j = f(\mathbf{x}_i;\Theta) + \mathbf{n}_i^j = \bar{\mathbf{y}}_i + \mathbf{n}_i^j,$$

$$\mathcal{L}(\Theta,\Sigma) = \mathcal{L}_{\text{pred}}(\Theta,\Sigma) + \lambda\mathcal{L}_{\text{noise}}(\Theta,\Sigma),$$

在显著性预测模块，直接使用已有的非监督方法的预测结果，与本方法的预测结果使用交叉熵损失。

$$L_{\text{CE}} = -(y\log(\hat{y}) + (1-y)\log(1-\hat{y})).$$

$$\mathcal{L}_{\text{pred}}(\Theta, \boldsymbol{\Sigma}) = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{m,n} L_{\text{CE}}(\mathbf{y}_{i,mn}^{j}, \hat{\mathbf{y}}_{i,mn}^{j}),$$

在噪声模块，假设噪声服从$N(0, \sigma)$。然后使用已有的非监督方法的预测结果建模$p(\sigma')$，然后迭代优化$q(\sigma)$.

$$\mathcal{L}_{\text{noise}}(\Theta, \Sigma) = \sum_{i}^{N} \mathbf{KL}(q(\boldsymbol{\Sigma}_i) \| \, p(\hat{\boldsymbol{\Sigma}}_i)).$$

$$\mathbf{KL}(q(\sigma) \| \, p(\hat{\sigma})) = \log(\hat{\sigma}/\sigma) + \frac{\sigma^2 + (\mu - \hat{\mu})^2}{2\hat{\sigma}^2} - \frac{1}{2},$$

$$(\sigma_i^{t+1})^2 = (\sigma_i^t)^2 + \alpha((\hat{\sigma}_i^t)^2 - (\sigma_i^t)^2),$$

训练策略：第一轮训练，对noise model进行零方差初始化，训练FCN到收敛。

Experimental Results

基于DeepLab network(ResNet-101 in particular)进行实验。在7个数据集上进行测试，同时与监督和非监督的SOTA方法进行了对比。值得一提的是，这里设计了3个baseline:

- *BL1: using noisy unsupervised saliency pseudo ground truth*
- *BL2: using averaged unsupervised saliency as pseudo ground truth*
- *BL3: supervised learning with ground truth supervision*

Table 1. Performance of mean F-measure (F_β) and MAE for different methods including ours on seven benchmark datasets.

	MSRA-B		ECSSD		DUT		SED2		PASCALS		THUR		SOD	
Methods	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
BL1	.7905	.0936	.7205	.1444	.5825	.1369	.7773	.1112	.6714	.2206	.5953	.1339	.6306	.1870
BL2	.6909	.1710	.6542	.2170	.4552	.2951	.7232	.1406	.6776	.2409	.5119	.2545	.5928	.2566
BL3	.8879	.0587	.8717	.0772	.7253	.0772	.8520	.0819	.8264	.1525	.7368	.0749	.7922	.1231
OURS	.8770	.0560	.8783	.0704	.7156	.0860	.8380	.0881	.8422	.1391	.7322	.0811	.7976	.1182

Table 2. Performance of mean F-measure (F_β) and MAE for different methods including ours on seven benchmark datasets (Best ones in bold). From DSS to DC are deep learning based supervised methods, from DRFI to HS are the handcrafted feature based unsupervised methods, SBF and OURS are deep learning based unsupervised saliency detection methods.

	MSRA-B		ECSSD		DUT		SED2		PASCALS		THUR		SOD	
Methods	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
DSS [11]	.8941	.0474	.8796	.0699	.7290	.0760	.8236	.1014	.8243	.1546	.7081	.1142	.8048	.1118
NLDF [26]	.8970	.0478	**.8908**	.0655	**.7360**	.0796	-	-	.8391	.1454	-	-	**.8235**	**.1030**
Amulet [40]	-	-	.8825	**.0607**	.6932	.0976	**.8745**	**.0629**	.8371	**.1292**	.7115	.0937	.7729	.1248
UCF [41]	-	-	.8521	.0797	.6595	.1321	.8444	.0742	.8060	.1492	.6920	.1119	.7429	.1527
SRM [29]	.8506	.0665	.8260	.0922	.6722	.0846	.7447	.1164	.7766	.1696	.6894	.0871	.7246	.1369
DMT [22]	-	-	.7589	.1601	.6045	**.0758**	.7778	.1074	.6657	.2103	.6254	.0854	.6978	.1503
RFCN [28]	-	-	.8426	.0973	.6918	.0945	.7616	.1140	.8064	.1662	.7062	.1003	.7531	.1394
DeepMC [42]	.8966	.0491	.8061	.1019	.6715	.0885	.7660	.1162	.7327	.1928	.6549	.1025	.6862	.1557
MDF [19]	.7780	.1040	.8097	.1081	.6768	.0916	.7658	.1171	.7425	.2069	.6670	.1029	.6377	.1669
DC [20]	**.8973**	**.0467**	.8315	.0906	.6902	.0971	.7840	.1014	.7861	.1614	.6940	.0959	.7603	.1208
DRFI [14]	.7282	.1229	.6440	.1719	.5525	.1496	.7252	.1373	.5745	.2556	.5613	.1471	.5440	.2046
RBD [43]	.7508	.1171	.6518	.1832	.5100	.2011	.7939	.1096	.6581	.2418	.5221	.1936	.5927	.2181
DSR [21]	.7227	.1207	.6387	.1742	.5583	.1374	.7053	.1452	.5785	.2600	.5498	.1408	.5500	.2133
MC [13]	.7165	.1441	.6114	.2037	.5289	.1863	.6619	.1848	.5742	.2719	.5149	.1838	.5332	.2435
HS [44]	.7129	.1609	.6234	.2283	.5205	.2274	.7168	.1869	.5948	.2860	.5157	.2178	.5383	.2729
SBF [35]	-	-	.7870	.0850	.5830	.1350	-	-	.7780	.1669	-	-	.6760	.1400
OURS	.8770	.0560	.8783	.0704	.7156	.0860	.8380	.0881	**.8422**	.1391	**.7322**	**.0811**	.7976	.1182

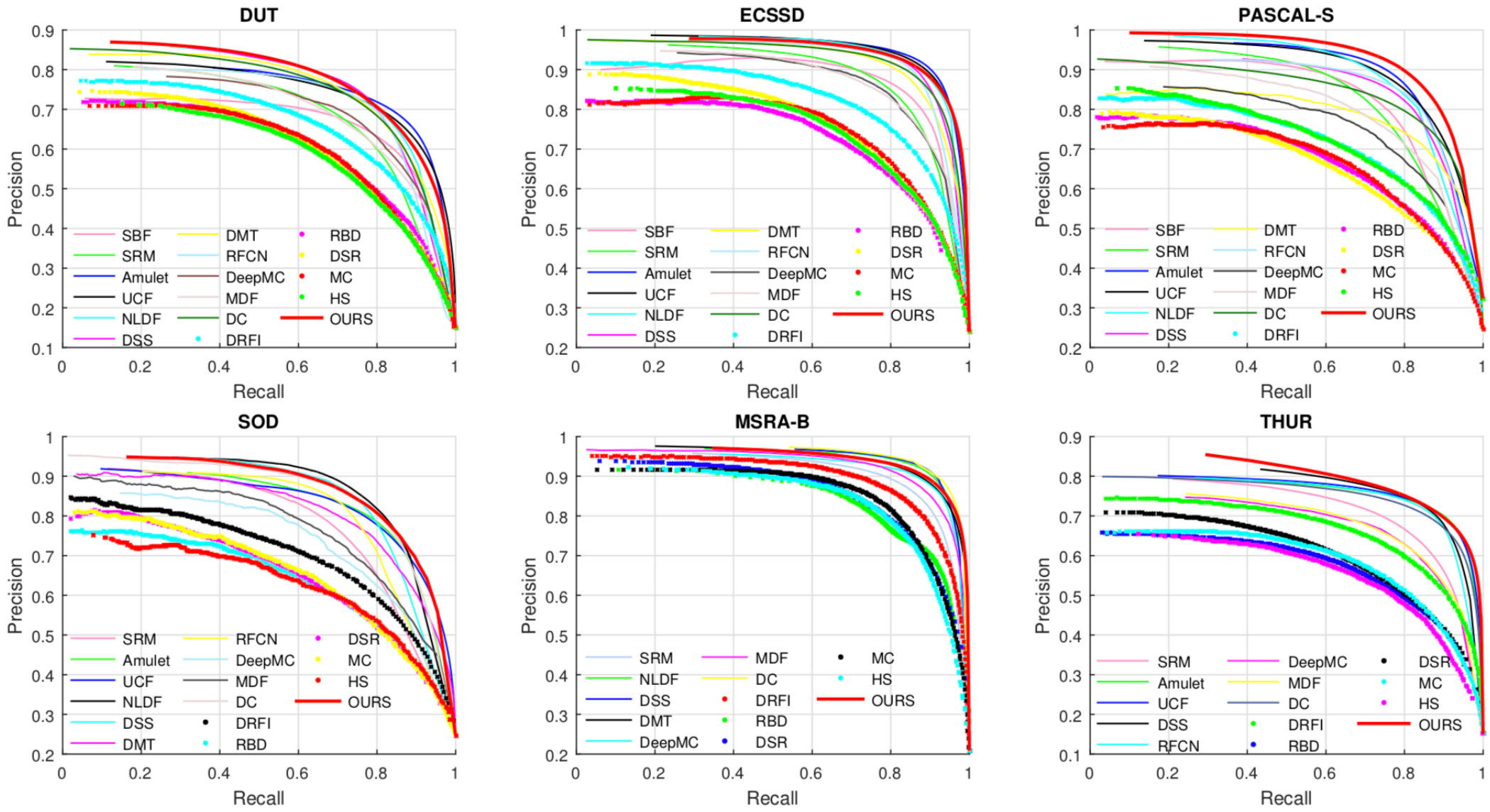

Figure 3. PR curves on six benchmark datasets (DUT, ECSSD, PASCAL-S, SOD, MSRA-B, THUR). Best Viewed on Screen.

Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection

这篇论文提出了一种基于新型递归网络架构的快速视频显著对象检测模型PDB-ConvLSTM，包括两个模块: PDC和PDB-ConvLSTM。
PDC模块首先被设计用于同时提取多尺度的空间特征。然后将这些空间特征连接起来并馈入扩展的更深层次的双向控制器DB-ConvLSTM以学习时空信息。
前向和后向ConvLSTM单元分为两层，以级联方式连接，鼓励双向流之间的信息流，从而导致更深的特征提取。本文通过采用几个扩展的DB-ConvLSTM来提取多尺度的时空信息，进一步增强了DB-ConvLSTM的PDC结构。
大量实验结果表明，本文的方法在很大范围内优于以前的视频显著性模型，单个GPU上的实时速度为20 fps。以unsupervised segmentation task作为示例应用，所提出的模型（具有基于CRF的后处理）在两个流行的基准上实现了SOTA结果，很好地展示了其优越的性能和高适用性。

Framework

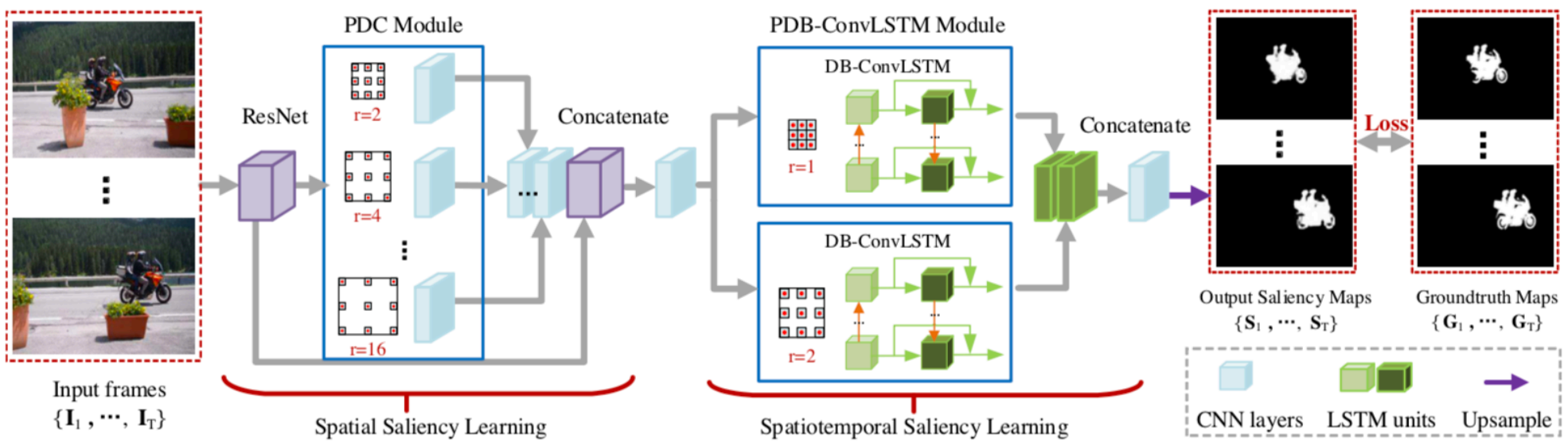

Fig. 1. Architecture overview of the proposed video salient object detection model, which consists of two components, *e.g.*, a spatial saliency learning module based on Pyramid Dilated Convolution (PDC) (§ 3.1) and a spatiotemporal saliency learning module via Pyramid Dilated Bidirectional ConvLSTM (PDB-ConvLSTM) (§ 3.2).

Spatial Saliency Learning via PDC Module

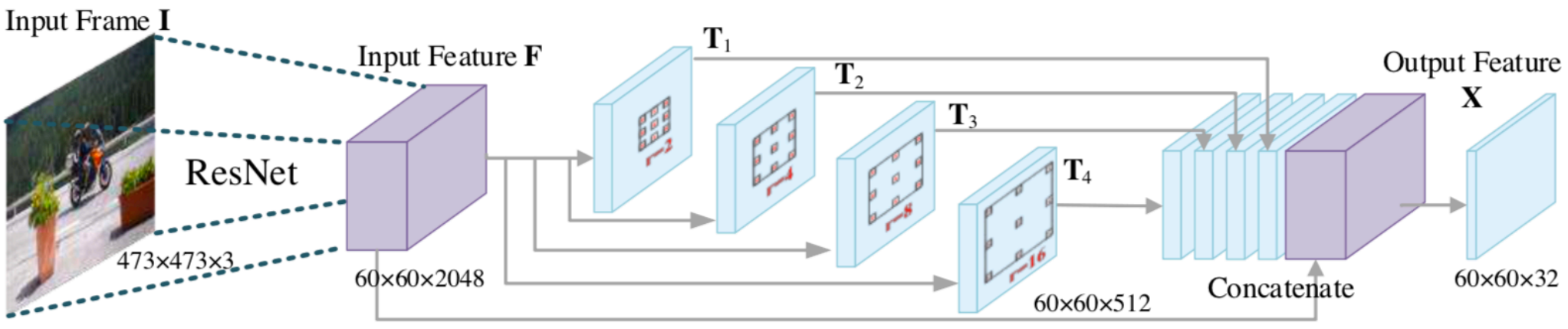

Fig. 2. Illustration of PDC module, where features from 4 parallel dilated convolution branches with different dilated rates are concatenated with the input features for emphasizing multi-scale spatial feature learning. See § 3.1 for details.

Spatiotemporal Saliency Learning via PDB-ConvLSTM Module

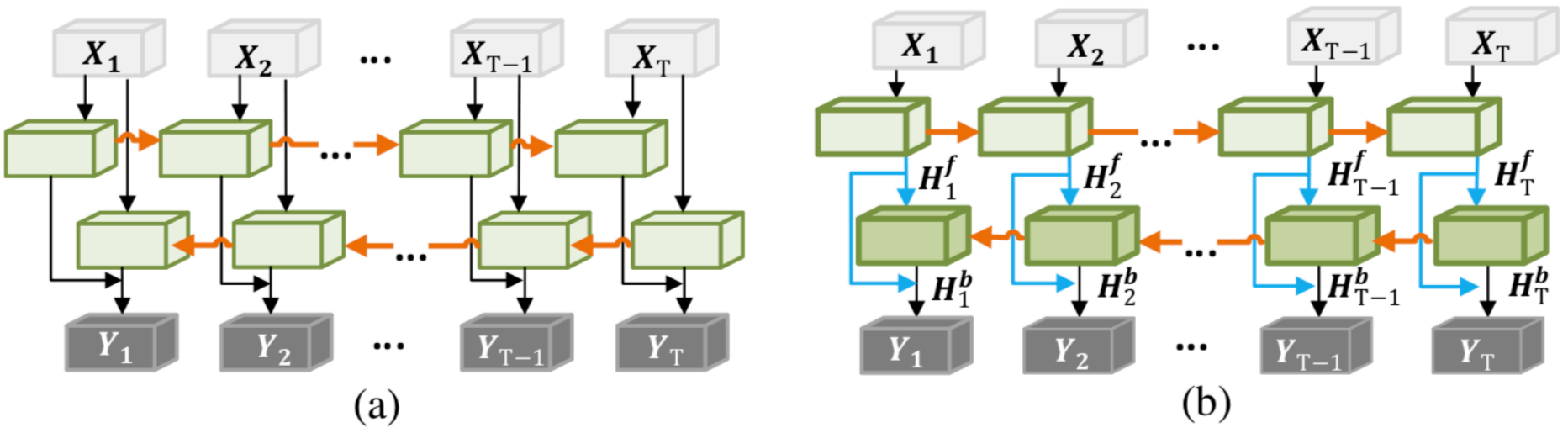

Fig. 3. Illustration of (a) Bidirectional ConvLSTM and (b) the proposed DB-ConvLSTM module. In PDB-ConvLSTM module, two DB-ConvLSTMs with different dilate rates are adopted for capturing multi-scale information and encouraging information flow between bi-directional LSTM units. See § 3.2 for details.

损失函数的设计

为了更有效地生成更好的saliency prediction和更有效地训练模型，本文提出了一个融合损失函数，它考虑了多个评估指标: cross entropy loss and MAE loss.

$$\mathcal{L}(\mathbf{S},\mathbf{G}) = \mathcal{L}_{cross\_entropy}(\mathbf{S},\mathbf{G}) + \mathcal{L}_{MAE}(\mathbf{S},\mathbf{G}),$$

$$\mathcal{L}_{cross\_entropy}(\mathbf{S},\mathbf{G}) = -\frac{1}{N}\sum_{i=1}^{N}\left[g_i log(s_i) + (1-g_i) log(1-s_i)\right]$$

$$\mathcal{L}_{MAE}(\mathbf{S},\mathbf{G}) = \frac{1}{N}\sum_{i=1}^{N}|g_i - s_i|.$$

Experimental Results

Performance on Video Salient Object Detection

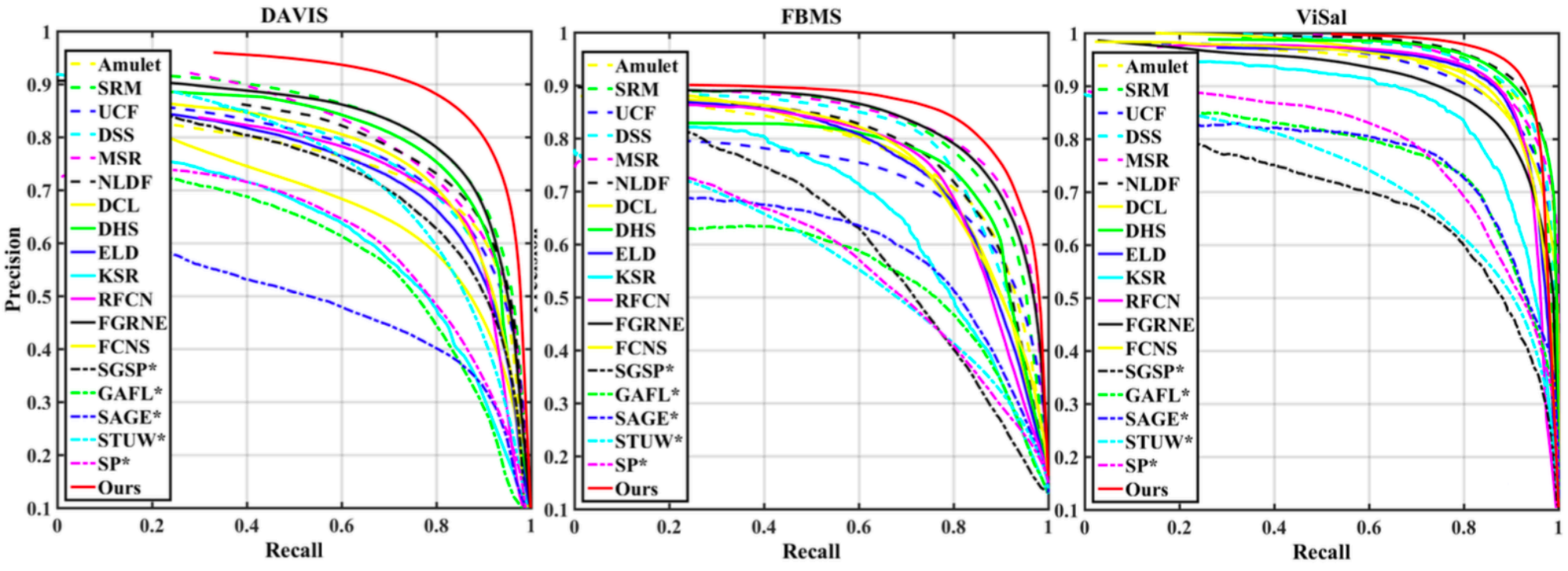

Fig. 4. Quantitative comparison against 18 saliency methods using PR curve on DAVIS [31], FBMS [2] and ViSal [43] datasets. Please see § 4.1 for more details.

Table 1. Quantitative comparison results against 18 saliency methods using MAE and maximum F-measure on DAVIS [31], FBMS [2] and ViSal [43]. The best scores are marked in **bold**. See § 4.1 for more details.

	Methods	Year	DAVIS		FBMS		ViSal	
			MAE↓	F^{max}↑	MAE↓	F^{max}↑	MAE↓	F^{max}↑
Image Saliency Models	Amulet [51]	ICCV'17	0.082	0.699	0.110	0.725	0.032	0.894
	SRM [36]	ICCV'17	0.039	0.779	0.071	0.776	0.028	0.890
	UCF [52]	ICCV'17	0.107	0.716	0.147	0.679	0.068	0.870
	DSS [16]	CVPR'17	0.062	0.717	0.083	0.764	0.028	0.906
	MSR [23]	CVPR'17	0.057	0.746	**0.064**	0.787	0.031	0.901
	NLDF [29]	CVPR'17	0.056	0.723	0.092	0.736	0.023	0.916
	DCL [25]	CVPR'16	0.070	0.631	0.089	0.726	0.035	0.869
	DHS [26]	CVPR'16	0.039	0.758	0.083	0.743	0.025	0.911
	ELD [22]	CVPR'16	0.070	0.688	0.103	0.719	0.038	0.890
	KSR [37]	ECCV'16	0.077	0.601	0.101	0.649	0.063	0.826
	RFCN [35]	ECCV'16	0.065	0.710	0.105	0.736	0.043	0.888
Video Saliency Models	FGRNE [24]	CVPR'18	0.043	0.786	0.083	0.779	0.040	0.850
	FCNS [44]	TIP'18	0.053	0.729	0.100	0.735	0.041	0.877
	SGSP* [27]	TCSVT'17	0.128	0.677	0.171	0.571	0.172	0.648
	GAFL* [43]	TIP'15	0.091	0.578	0.150	0.551	0.099	0.726
	SAGE* [42]	CVPR'15	0.105	0.479	0.142	0.581	0.096	0.734
	STUW* [8]	TIP'14	0.098	0.692	0.143	0.528	0.132	0.671
	SP* [28]	TCSVT'14	0.130	0.601	0.161	0.538	0.126	0.731
	Ours	ECCV'18	**0.030**	**0.849**	0.069	**0.815**	**0.022**	**0.917**

* Non-deep learning model.

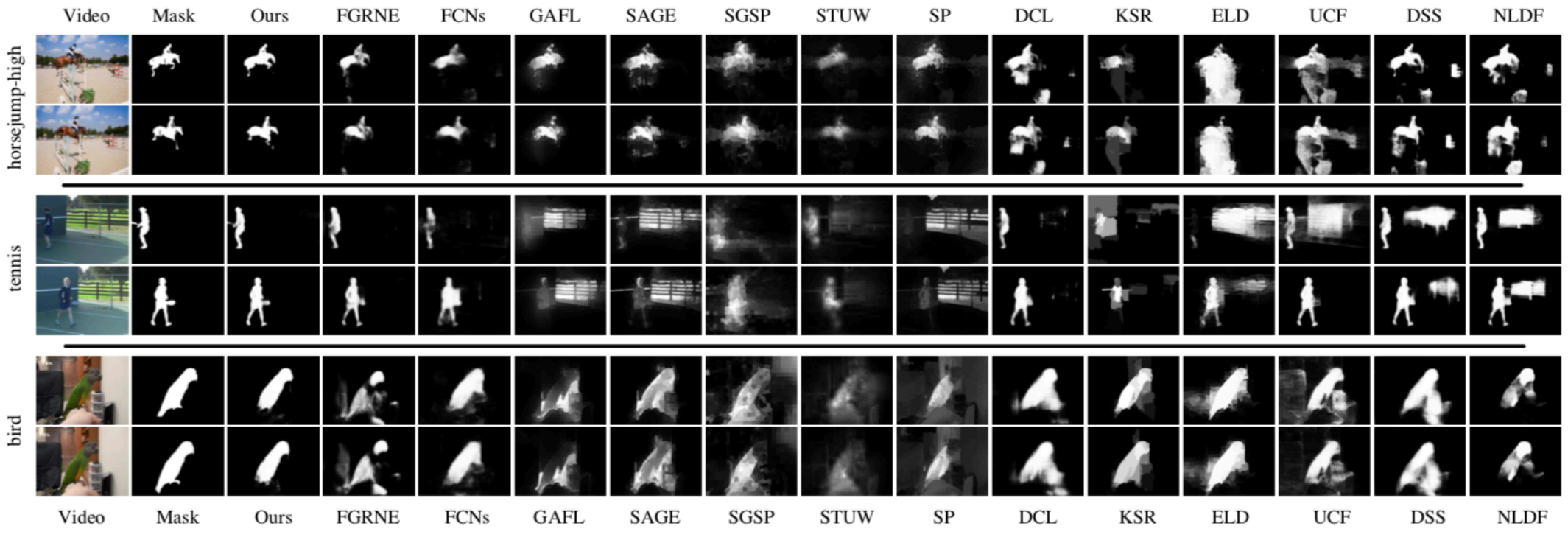

Fig. 5. Qualitative comparison against other top-performing saliency methods with groundtruths on three example video sequences. Zoom-in for details.

Performance on Unsupervised Video Object Segmentation

Table 2. Comparison with 7 representative unsupervised video object segmentation methods on the test sets of DAVIS and FBMS datasets. The best scores are marked in **bold**. See § 4.2 for details.

Dataset	Metric	Method								
		ARP*[20]	LVO[34]	FSEG[19]	LMP[33]	SFL*[5]	FST*[30]	SAGE*[42]	Ours	Ours+
DAVIS	$\mathcal{J}\uparrow$	76.2	75.9	70.7	70.0	67.4	55.8	41.5	74.3	**77.2**
	$\mathcal{F}\uparrow$	70.6	72.1	65.3	65.9	66.7	51.1	36.9	72.8	**74.5**
FBMS	$\mathcal{J}\uparrow$	59.8	65.1	68.4	35.7	55.0	47.7	61.2	72.3	**74.0**

* Non-deep learning model.

Runtime Analysis

Table 3. Runtime comparison with 6 existing video saliency methods.

Method	SGSP[27]	SAGE[42]	GAFL[43]	STUW[8]	SP[28]	FCNS[44]	Ours
Time(s)	1.70*(+)	0.88*(+)	1.04*(+)	0.78*(+)	6.05*(+)	0.47	**0.05**

* CPU time.
(+) indicates extra computation of optical flow. For reference, LDOF [1] takes about 49.64s per frame, Flownet v2.0 [17] takes about 0.05s per frame.

Unsupervised CNN-based Co-Saliency Detection with Graphical Optimization

这篇论文通过Unsupervised CNN解决了一组共同覆盖特定类对象的图像中的co-saliency detection。尽管监督方法可以实现feature learning和co-saliency detection的整合，但是它们需要以object masks形式提供额外的训练数据，通过手动绘制或由具有密集用户交互的工具描绘。此外，这些学习模型在测试中对于unseen object categories可能表现不佳，因为模型不能适应unseen categories。
而本文方法不需要以object masks形式的任何额外训练数据。本文将co-saliency detection分解为两个子任务，single-image saliency detection和cross-image co-occurrence region discovery，对应于两个新的无监督损失，single-image saliency (SIS) loss和co-occurrence (COOC) loss。这两个损失是在图形模型上建模的，其中前者和后者分别作为unary和pairwise。可以联合优化这两个任务以生成高质量的co-saliency maps。
此外，可以通过两个扩展来增强所生成的co-saliency maps的质量：map sharpening by self-paced learning和boundary preserving by fully connected conditional random fields。实验表明，本文的方法取得了优异的成果，甚至超过了许多监督方法。

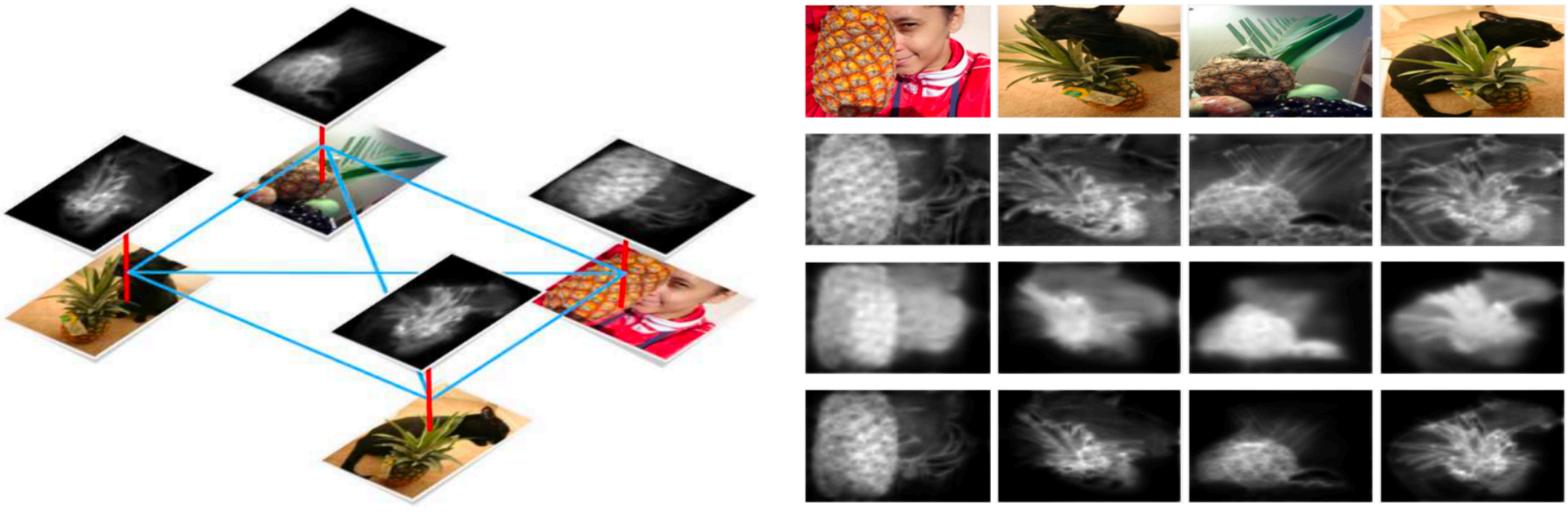

Fig. 1. Motivation of our method. (a) Our method optimizes an objective function defined on a graph where single-image saliency (SIS) detection (red edges) and cross-image co-occurrence (COOC) discovery (blue edges) are considered jointly. (b) The first row displays the images for co-saliency detection. The following three rows show the detected saliency maps by using COOC, SIS, and both of them, respectively.

Framework

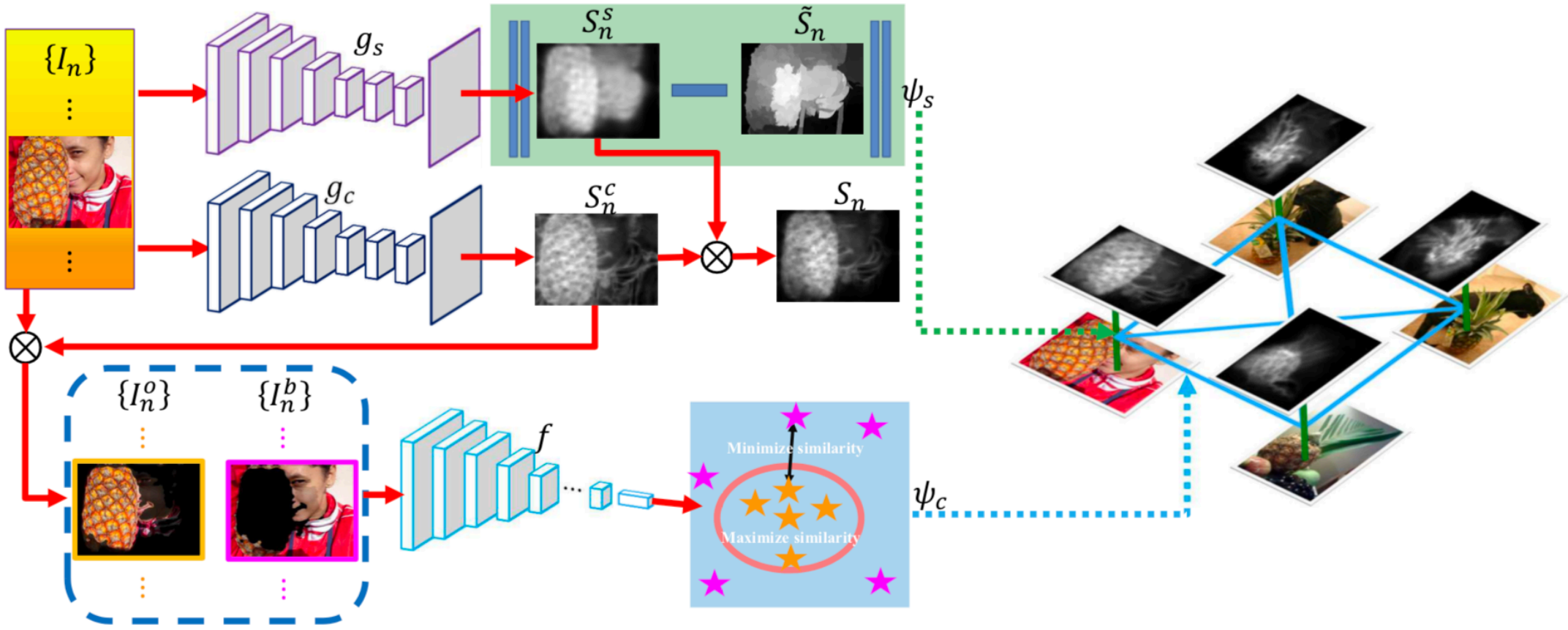

Fig. 2. Overview of our approach to co-saliency detection. It optimizes an objective function defined on a graph by learning two collaborative FCN models g_s and g_c which respectively generates single-image saliency maps and cross-image co-occurrence maps.

The proposed objective function on the graph is defined by

$$E(\mathbf{w}) = \sum_{n=1}^{N} \psi_s(I_n; \mathbf{w}) + \sum_{n=1}^{N} \sum_{m \neq n} \psi_c(I_n, I_m; \mathbf{w}),$$

Unary term ψs

$$\psi_s(I_n;\mathbf{w}_s)=\sum_{i\in I_n} R_n(i)|S_n^s(i)-\tilde{S}_n(i)|^2,$$

Pairwise term ψc

$$I_n^o=S_n^c\otimes I_n \quad \text{and} \quad I_n^b=(1-S_n^c)\otimes I_n,$$

$$\psi_c(I_n,I_m;\mathbf{w}_c)=-\log(p_{nm}),$$

$$p_{nm}=\frac{\exp(-d_{nm}^+)}{\exp(-d_{nm}^+)+\exp(-d_{nm}^-)}, \text{ where}$$

$$d_{nm}^+=\frac{1}{c}\|f(I_n^o)-f(I_m^o)\|^2 \text{ and}$$

$$d_{nm}^-=\frac{1}{2c}\|f(I_n^o)-f(I_n^b)\|^2+\frac{1}{2c}\|f(I_m^o)-f(I_m^b)\|^2.$$

$$\psi_c(I_n,I_m;\mathbf{w}_c)=-\lambda_c\log(p_{nm})-\lambda_{\tilde{c}}\log(\tilde{p}_{nm}),$$

Co-saliency map enhancement

The self–paced learning with CNNs is proposed to make salience map sharper. Then, fully connected conditional random fields are adopted to preserve co–salient objects’ boundaries.

- Co–saliency map enhancement by self–paced learning.
- Postprocessing using DenseCRFs.

$$q_n^k\in\begin{cases}O_n, & \text{if } \mu_n^k>\mu_n+\sigma_n,\\ B_n, & \text{if } \mu_n^k<\mu_n-0.25*\sigma_n, \text{ for } k=1,2,...,K,\\ T_n, & \text{otherwise,}\end{cases}$$

$$\psi_e(I_n;\mathbf{w}_e)=w_o\sum_{q\in O_n}\sum_{i\in q}|S_n^e(i)-1|^2+w_b\sum_{q\in B_n}\sum_{i\in q}|S_n^e(i)-0|^2,$$

$$E(\mathbf{w}) = \sum_{n=1}^{N} \psi_s(I_n; \mathbf{w}_s) + \lambda_e \psi_e(I_n; \mathbf{w}_e) + \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} \psi_c(I_n, I_m; \mathbf{w}_c),$$

Experimental Results

Table 1. The performance of co-saliency detection on three benchmark datasets. SI and CS denote the single-image saliency and co-saliency methods, respectively. US and S indicate the unsupervised and supervised methods, respectively. The numbers in red and green respectively indicate the best and the second best results of the unsupervised co-saliency methods (CS+US), the group which the proposed method belongs to.

Method	Setting	MSRC			iCoseg			Cosal2015		
		AP	F_β	S_α	AP	F_β	S_α	AP	F_β	S_α
DIM [17]	CS+S	-	-	-	0.8773	0.7918	0.7583	-	-	-
UMLBF [13]	CS+S	0.9160	0.8410	-	-	-	-	0.8210	0.7120	-
CBCS [7]	CS+US	0.7034	0.5910	0.4801	0.7972	0.7408	0.6580	0.5863	0.5579	0.5439
SACS [31]	CS+US	0.8602	0.7877	0.7074	0.8400	0.7973	0.7523	0.7077	0.6923	0.6938
CSHS [8]	CS+US	0.7834	0.7118	0.6661	0.8454	0.7549	0.7502	0.6198	0.6181	0.5909
ESMG [32]	CS+US	0.6659	0.6245	0.5804	0.8347	0.7766	0.7677	0.5133	0.5114	0.5446
CSSCF [3]	CS+US	0.8604	0.8005	0.7383	0.8400	0.7811	0.7404	0.7075	0.6815	0.6710
CoDW [12]	CS+US	0.8435	0.7724	0.7129	0.8766	0.7985	0.7500	0.7438	0.7046	0.6473
SP-MIL [11]	CS+US	0.8974	0.8029	0.7687	0.8749	0.8143	0.7715	-	-	-
MVSRC [55]	CS+US	0.8530	0.7840	-	0.8680	0.8100	-	-	-	-
Ours	CS+US	0.9226	0.8404	0.7948	0.9112	0.8497	0.8200	0.8149	0.7580	0.7506
LEGS [26]	SI+S	0.8479	0.7701	0.6997	0.7924	0.7473	0.7529	0.7339	0.6926	0.7068
DCL [47]	SI+S	0.9065	0.8259	0.7742	0.9003	0.8444	0.8606	0.7815	0.7386	0.7591
DSS [28]	SI+S	0.8700	0.8313	0.7435	0.8802	0.8386	0.8483	0.7745	0.7509	0.7579
UCF [29]	SI+S	0.9217	0.8114	0.8175	0.9292	0.8261	0.8754	0.8081	0.7194	0.7790
Amulet [30]	SI+S	0.9219	0.8159	0.8162	0.9395	0.8381	0.8937	0.8201	0.7384	0.7856
GMR [20]	SI+US	0.8092	0.7460	0.6547	0.7990	0.7805	0.7068	0.6649	0.6605	0.6599
GP [22]	SI+US	0.8200	0.7422	0.6844	0.7821	0.7495	0.7198	0.6847	0.6576	0.6714
MB+ [23]	SI+US	0.8367	0.7817	0.7200	0.7868	0.7706	0.7272	0.6710	0.6689	0.6724
MST [24]	SI+US	0.8057	0.7491	0.6460	0.8019	0.7659	0.7292	0.7096	0.6669	0.6676
MILP [25]	SI+US	0.8334	0.7776	0.6871	0.8182	0.7883	0.7514	0.6797	0.6734	0.6752
SVFSal [42]	SI+US	0.8669	0.7934	0.7688	0.8376	0.8056	0.8271	0.7468	0.7120	0.7604

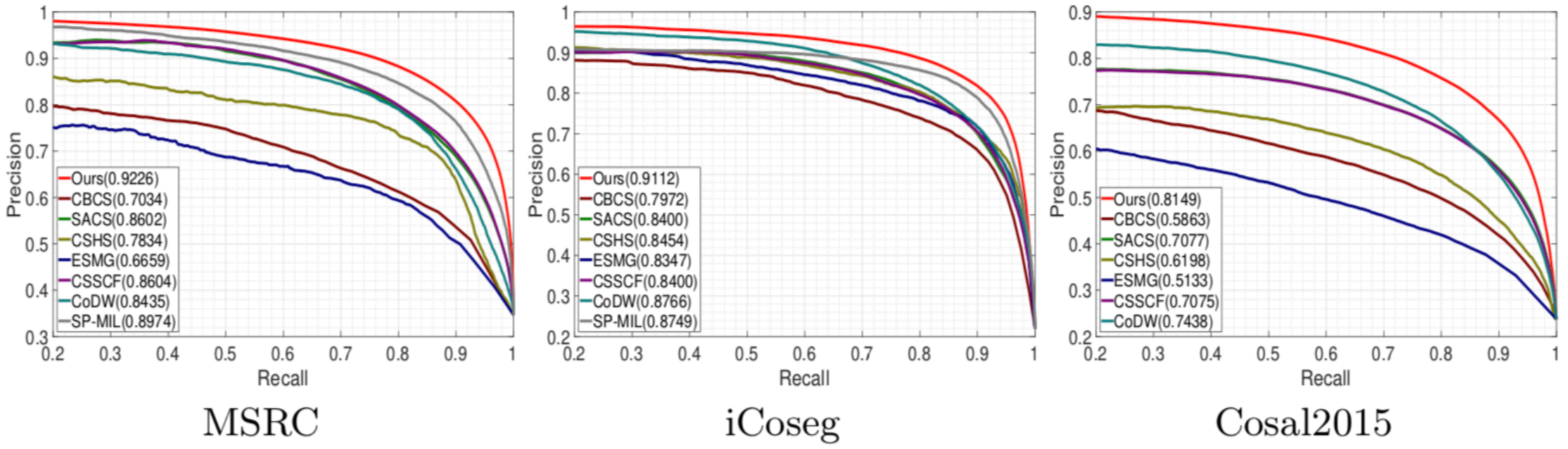

Fig. 3. Comparison with the state-of-the-art methods with the same setting in terms of PR curves on three benchmark datasets. The numbers in parentheses are AP values.

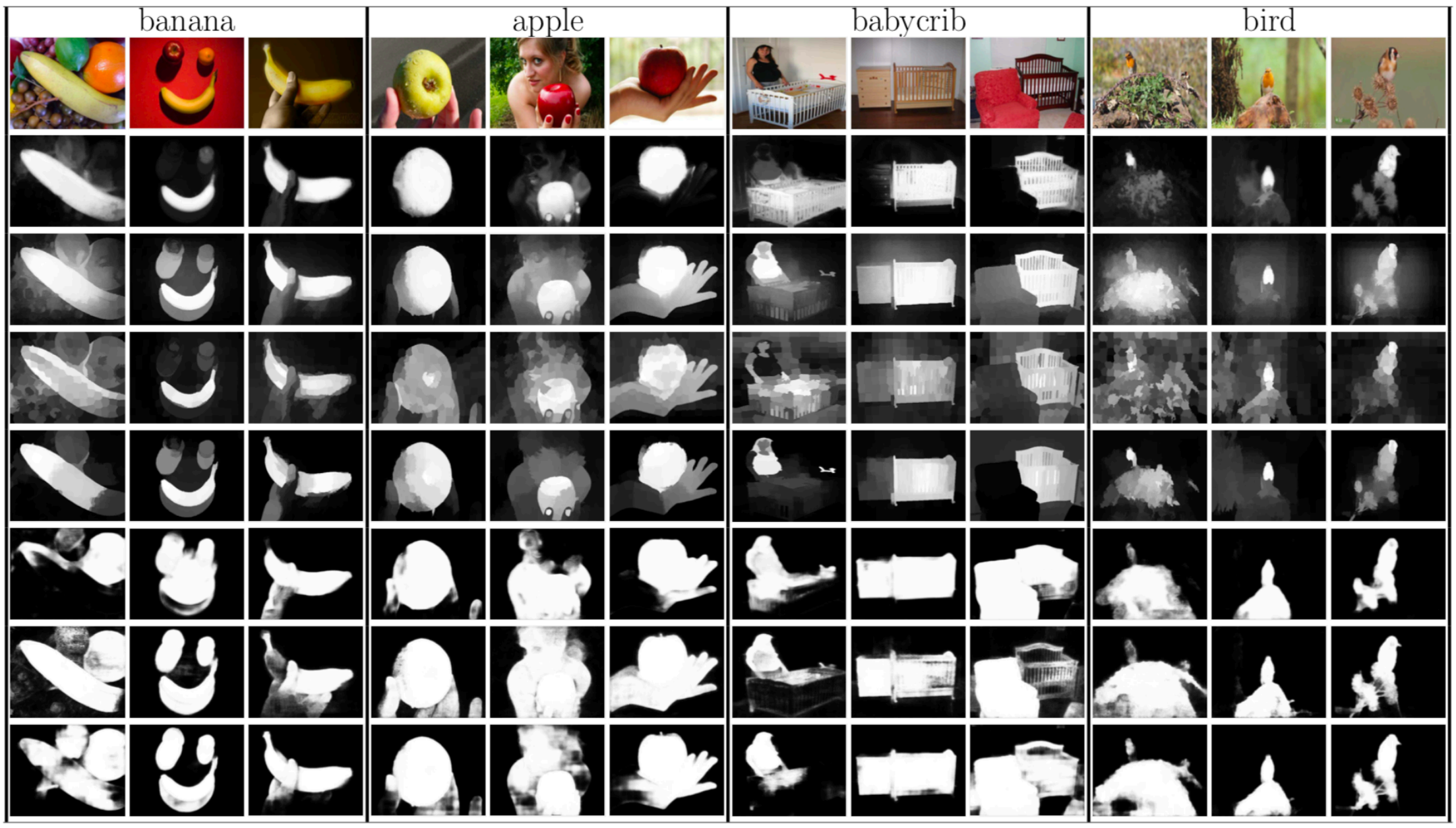

Fig. 4. Example saliency maps generated by our method and some state-of-the-art methods. From the top to the bottom, they are the given images, ours, `CSSCF` [3], `CoDW` [12], `MILP` [25], `SVFSal` [42], `UCF` [29] and `Amulet` [30].

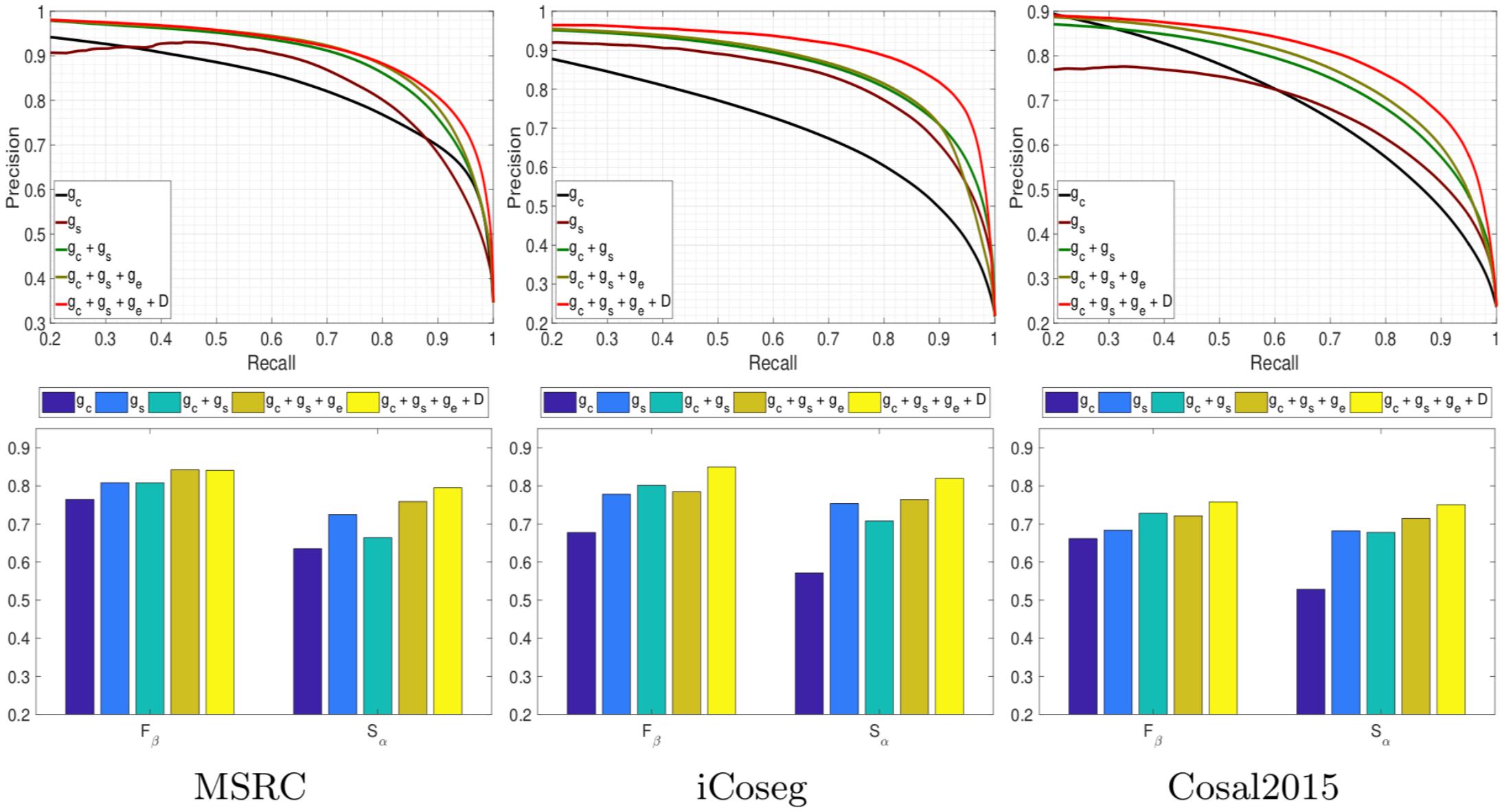

Fig. 5. Ablation studies on three benchmarks. The top row plots the PR curves, while the bottom row shows the performance in F_β and S_α.

Fig. 6. Example co-saliency maps generated by combinations of different components. From the top to the bottom, they are the given images, g_c, g_s, g_c+g_s, g_c+g_s+g_e and g_c+g_s+g_e+D, respectively.