Saliency Detection by Fully Learning a Continuous Conditional Random Field

Keren Fu, Irene Yu-Hua Gu, *Senior Member, IEEE*, and Jie Yang

***Abstract*—Salient object detection is aimed at detecting and segmenting objects that human eyes are most focused on when viewing a scene. Recently, conditional random field (CRF) is drawn renewed interest, and is exploited in this field. However, when utilizing a CRF with unary and pairwise potentials having essential parameters, most existing methods only employ manually designed parameters, or learn parameters partly for the unary potentials. Observing that the saliency estimation is a continuous labeling issue, this paper proposes a novel data-driven scheme based on a special CRF framework, the so-called continuous CRF (C-CRF), where parameters for both unary and pairwise potentials are jointly learned. The proposed C-CRF learning provides an optimal way to integrate various unary saliency features with pairwise cues to discover salient objects. To the best of our knowledge, the proposed scheme is the first to completely learn a C-CRF for saliency detection. In addition, we propose a novel formulation of pairwise potentials that enables learning weights for different spatial ranges on a superpixel graph. The proposed C-CRF learning-based saliency model is tested on 6 benchmark datasets and compared with 11 existing methods. Our results and comparisons have provided further support to the proposed method in terms of precision-recall and F-measure. Furthermore, incorporating existing saliency models with pairwise cues through the C-CRF are shown to provide marked boosting performance over individual models.**

***Index Terms*—Continuous conditional random field (C-CRF), feature integration, learning, saliency map, salient object detection, spatial ranges.**

I. Introduction

SALIENCY detection is aimed at detecting conspicuous image parts that attract human attention. It simulates and models the selective mechanism of human eyes [1], [2]. There are generally two subcategories of saliency detection: eye-fixation prediction [3]–[6] and salient object/region detection [7]–[10]. The former task aims to detect sparse eye-fixation points where human attend in a scene, whereas the latter task is to detect and emphasize entire salient objects from an image, yielding a *saliency map* as output where the pixel-wise intensities indicate the probability of being a salient object. The recent advance in salient object detection is driven by emerging multimedia applications such as automatic object detection and segmentation [11]–[13], content-based image editing [14]–[18], image retrieval [19]–[21] and compression, image sequence and video analysis [22], [23]. In this paper, we mainly address salient object detection.

To emphasize salient objects uniformly, the conditional random field (CRF) that can provide label consistency becomes popular in this field. By utilizing CRF, high quality saliency maps that maintain well-defined object boundaries and uniformly emphasized object interior are achieved. In existing studies [24]–[27], CRFs are employed in explicit or implicit ways. However, when utilizing a CRF whose energy function consists of parameterized unary and pairwise energy potentials, most previous methods use manually designed parameters [25] or learn the parameters only for unary potentials [24], [27]. Hence for saliency detection, the full power of CRF on feature integration is hardly exploited.

Motivated by the above issues, this paper proposes to fully learn a CRF, namely to learn both unary and pairwise parameters in order to exploit the power of CRF for feature integration in saliency detection [1], [3]. More specifically, we investigate a special CRF framework—*continuous CRF (C-CRF)* [28]–[30]. This is motivated by the idea that saliency detection is conventionally treated as a *continuous labeling problem*. In this paper a novel data-driven saliency detection scheme based on C-CRF [28] is proposed, which differs from [24], [27] since ours enables learning to integrate various pairwise features. This allows the C-CRF model to capture more sophisticated interactions between image parts, leading to enhanced delineation between objects and background in the resulting saliency maps. It is worth noting that the work of Mai *et al.* [26] is closely related to ours. In [26], the unary and pairwise potentials of CRF all include parameters. However, the main difference in between is that [26] employs discrete CRF (as will be mentioned later), whereas we propose to use C-CRF which benefits from different designs of energy function, hence very different techniques for learning and inference. In addition, as shown in Section V-C the proposed method improves the performance significantly from [26]. A straightforward comparison of the

Manuscript received September 5, 2016; revised January 23, 2017; accepted February 23, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61572315 and Grant 6151101179, and in part by the 863 Plan of China under Grant 2015AA042308. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Judith Redi. *(Corresponding authors: Jie Yang and Keren Fu.)*

K. Fu is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Department of Signals and Systems, Chalmers University of Technology, Gothenburg 41296, Sweden (e-mail: fkrshichaoren@qq.com).

I. Y.-H. Gu is with the Department of Signals and Systems, Chalmers University of Technology, Gothenburg 41296, Sweden (e-mail: irenegu@chalmers.se).

J. Yang is with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: jieyang@sjtu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2017.2679898

TABLE I
COMPARISON OF REPRESENTATIVE CRF-BASED METHODS IN THE SALIENT OBJECT/REGION DETECTION COMMUNITY

Related work	CRF type	Learning unary terms	Learning pairwise terms
Liu *et al.* [24]	Discrete (D-CRF)	Yes	No
Mai *et al.* [26]	Discrete (D-CRF)	Yes	Yes
Yang *et al.* [25]	Continuous (C-CRF)	No	No
Lu *et al.* [27]	Continuous (C-CRF)	Yes	No
Ours	**Continuous (C-CRF)**	**Yes**	**Yes**

The abbreviation C-CRF and D-CRF stand for continuous CRF and discrete CRF, respectively.

proposed method to state-of-the-art CRF related works are given in Table I. To the best of our knowledge, the complete C-CRF learning and inference theories have not yet been applied to saliency estimation.

C-CRF was firstly proposed for ranking documents [28], and later applied to recognition [29] and depth estimation [30]. It is worth noting that CRF has already been applied to figure-ground segmentation [31], semantic segmentation [31]–[33], and also saliency detection [24], [26], [27] (Table I). However, most of them are conventional CRFs with *discrete* labels. We will later call this type of CRFs as D-CRF (discrete CRF). In the context of saliency detection, C-CRF may suit this problem better since saliency maps are known to be continuous and real-valued [3], [8], [34], revealing saliency detection can be regarded as a continuous labeling problem.

The main contributions of this paper are four-fold:

1) This study is the first to apply the complete C-CRF learning and inferring theories to saliency detection, leading to a data-driven way for saliency feature integration.
2) As shown in Table I, our work differs from existing saliency models that have explicit/implict relation to CRF, evolving from partially learning unary terms [24], [27] to jointly learning both unary plus pairwise terms, and from discrete field [26] to continuous field.
3) We propose a novel formulation of pairwise potentials for C-CRF defined on a superpixel graph. Such a formulation is conducted by graph topology decomposition and enables learning pairwise parameters for different spatial ranges of graph connections. This avoids the manual effort of tuning spatial connections of a graph.
4) We show from tests and comparisons that integrating widely employed unary saliency features with pairwise cues in a C-CRF manner outperforms a range of state-of-the-art methods. Furthermore, integrating several best-performing state-of-the-art methods through a C-CRF further pushes the performance to a new high level.

The reminder of this paper is organized as follows. Section II briefly reviews the fundamental theories of CRF and C-CRF. Section III describes the related work. Section IV describes the proposed method. Experimental results, performance evaluation and comparisons are included in Section V. Finally, conclusion is drawn in Section VI.

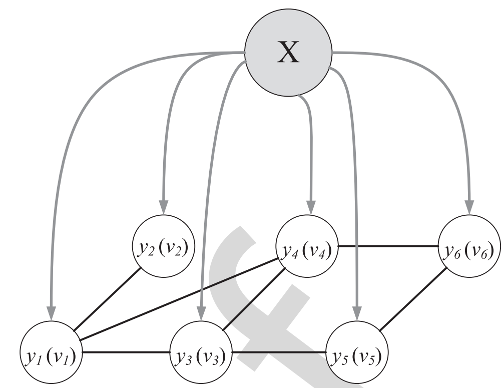

Fig. 1. General graphic model of CRF for image labeling task. A white vertex (v_i) represents a label (y_i) and the gray vertex ($\mathbf{x}$) represents the entire image. The gray arrows indicate the unary dependencies (conditions) while the black lines indicate the pairwise relations associating with a graph.

II. CONDITIONAL RANDOM FIELD (CRF) AND CONTINUOUS CONDITIONAL RANDOM FIELD (C-CRF): A BRIEF REVIEW

A. *Probabilistic Formulation*

Conditional random field (CRF) is originally proposed by Lafferty *et al.* [35] for labeling sequence data. For the image labeling task, given an image $\mathbf{x}$, the conditional probability distribution of a label configuration $\mathbf{y}$ (in vector form) on the CRF can be defined as

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\exp\{-\mathbb{E}(\mathbf{y},\mathbf{x})\} \tag{1}$$

where $\mathbb{E}(\mathbf{y},\mathbf{x})$ is the energy function and $Z(\mathbf{x})$ is the partition function.[1] The energy function can be expressed as unary terms plus pairwise terms as

$$\mathbb{E}(\mathbf{y},\mathbf{x}) = \sum_i \underbrace{U_{\boldsymbol{\alpha}}(y_i,\mathbf{x})}_{\text{Unary term}} + \sum_{i,j,i\sim j} \underbrace{P_{\boldsymbol{\varphi}}(y_i,y_j,\mathbf{x})}_{\text{Pairwise term}} \tag{2}$$

where y_i is the ith element of the label vector $\mathbf{y}$, $U_{\boldsymbol{\alpha}}$ and $P_{\boldsymbol{\varphi}}$ denote the unary and pairwise terms parameterized by vector $\boldsymbol{\alpha}$ and $\boldsymbol{\varphi}$ (vector $\boldsymbol{\alpha}$ contains the parameters for unary potentials, and vector $\boldsymbol{\varphi}$ contains the parameters for pairwise potentials). A CRF is often coupled with the definition of an undirected graph $G(V,E)$ [35], where V is the set of graph nodes and E is the set of graph edges. The label assigned to each graph node $v_i \in V$ is denoted as y_i. In (2), the notation "$i \sim j$" means that v_i and v_j are graph neighbors. Theoretically, the unary term $U_{\boldsymbol{\alpha}}$ represents the dependency between a label and the image $\mathbf{x}$ at a specific node, whereas the pairwise term $P_{\boldsymbol{\varphi}}$ encourages neighboring graph nodes to take similar labels (i.e., enforces labeling consistency). A general graphic model of CRF for image labeling task is given in Fig. 1, where a white vertex represents a label and the gray vertex represents the entire image.

[1] The partition functions for D-CRF and C-CRF are defined as $Z(\mathbf{x}) = \sum_{\mathbf{y}}\exp\{-\mathbb{E}(\mathbf{y},\mathbf{x})\}$ and $Z(\mathbf{x}) = \int_{\mathbf{y}}\exp\{-\mathbb{E}(\mathbf{y},\mathbf{x})\}d\mathbf{y}$, respectively.

B. D-CRF and C-CRF

In the conventional CRF, i.e., the D-CRF [31]–[33], [35], all components of $\mathbf{y}$ range over a finite label alphabet (e.g., subject to $\mathbf{y} \in \{0,1\}^n$ for a binary labeling problem, where n is the dimension of $\mathbf{y}$), whereas in the continuous CRF (C-CRF) [28], $\mathbf{y}$ is relaxed to be continuous values ($\mathbf{y} \in \mathcal{R}^n$). Due to such relaxing, the designs of energy functions for D-CRF and C-CRF differ. For example, D-CRF usually employs Potts model [32], [33] with indicator function for the pairwise terms, whereas in C-CRF, quadratic cost function can be used to measure the label compatibility. Besides, the techniques for learning and inference of C-CRF [28] differ significantly from D-CRF. The exact learning/inference of D-CRF is usually intractable due to its discrete property, which requires approximation techniques [36] such as belief propagation, mean field, Monte Carlo approaches, to name a few. In contrast, C-CRF offers direct learning together with closed-form inference, which will be shown later in this paper.

Assuming that the parameters of a CRF are given or estimated by learning, theoretically the optimal labeling vector $\mathbf{y}$ can be inferred by maximizing (1), or equivalently minimizing the negative logarithm of (1) as

$$-\log p(\mathbf{y}|\mathbf{x}) = \mathbb{E}(\mathbf{y},\mathbf{x}) + \log Z(\mathbf{x}). \tag{3}$$

Since $\log Z(\mathbf{x})$ is a constant with respect to $\mathbf{y}$, one can directly minimize the energy function $\mathbb{E}(\mathbf{y},\mathbf{x})$. From this viewpoint, existing methods on saliency detection such as manifold ranking [25], graph regularization [37], quadratic model [27] that minimize an energy function in the form of (2) can be viewed as special cases of inferring C-CRFs.

III. Related Work

A large number of literatures on salient object detection exist, see the comprehensive survey [38] and benchmarking [10]. Here we review some previous works that are highly relevant to data-driven approaches or CRF-based approaches.

Data driven approaches: The concept of *learning to detect* in saliency detection originates from [24], [39]. The idea behind is to automatically discover feature integration rules from training data instead of using manually designed rules. Judd *et al.* [39] propose to learn a saliency model from eye-tracking data, where low-, middle- and high-level image features are integrated by a linear SVM. Their work is, however, focused on eye-fixation prediction. Alex *et al.* [40] learn to score windows sampled from a given image, where the Bayesian theory is adopted for cue integration. The posterior constitutes the final objectness score of a window. Khuwuthyakorn *et al.* [41] learn to integrate pixel-wise saliency features via a mixture of linear SVMs. Mehrani *et al.* [42] use confidence scores from a boosting classifier to formulate a saliency map. After that, the saliency map is fed to a graph cut program for figure-ground segmentation. Jiang *et al.* [43] propose to extract abundant discriminative features from image regions. A random forest regressor is trained to map regional features to final saliency scores. Online saliency learning is proposed in [44], [45], where multiple kernel boosting is employed to identify salient parts against non-salient parts. Some recent data-driven methods [46], [47] consider deep learning for saliency detection. Due to the deep architecture of convolutional neural networks (CNNs), impressive performance is obtained. However, in CNNs there often lacks explicit modeling of neighborhood relations. Therefore, post-processing like C-CRF may be required.

CRF inference-based approaches: Several methods [25], [37] are based on inferring C-CRF without learning, where features and integration rules are manually specified. In [25], [37], though the word "CRF" or "continuous CRF" is not explicitly mentioned, there is a potential connection between these methods and C-CRF. To be more specific, the employed manifold ranking [25] and graph regularization [37] are special cases of inferring C-CRFs, as aforementioned in Section II-B.

CRF learning-based approaches: Some methods on saliency detection are based on both learning and inferring D-CRFs or C-CRFs. Learning is first conducted to obtain optimal parameters and inference is then applied on user-input images to achieve final saliency maps. Representative works include: Liu *et al.* [24] detect and segment salient objects by aggregating pixel saliency cues in a D-CRF. Linear weights for those cues are learned under the maximized likelihood (ML) criteria by tree-reweighted belief propagation. Mai *et al.* [26] propose a saliency aggregation approach, which aggregates saliency maps output by existing saliency models using a D-CRF. Weights for aggregation are learned from images retrieved from a predefined dataset. Lu *et al.* [27] learn optimal combination of seeds for graph-based diffusion by maximizing figure-ground segregation, where the employed graph diffusion is tightly related to C-CRF. The method boils down to learning the linear parameters of unary terms of the C-CRF. In summary, [24], [26] concern D-CRFs for saliency detection, where only unary parameters are learned. [27] implicitly considers a C-CRF, where again only the unary terms are learned. In contrast, our data-driven scheme differs from all the above methods on learning a complete C-CRF.

IV. The Proposed Method

This section describes the proposed method for saliency detection that is based on fully learning and inferring a continuous CRF (C-CRF). The block diagram of the proposed method is given in Fig. 2. An input image is first over-segmented into superpixels and a superpixel graph is established to capture intrinsic image context. A C-CRF will later be defined in conjunction with this graph. Next, we extract various unary saliency features and pairwise cues, which will be used to compose the unary and pairwise terms in the C-CRF energy function. By utilizing the off-line learned C-CRF parameters for both unary and pairwise potentials, the inference of the C-CRF corresponds to a final saliency map that is continuously valued. Details of each part of the method are further given in the following subsections.

A. Graph Construction From an Image

We first describe the graph construction, where the C-CRF is defined upon. Rather than constructing CRF on the pixel level [24], the proposed C-CRF is constructed on superpixels, where

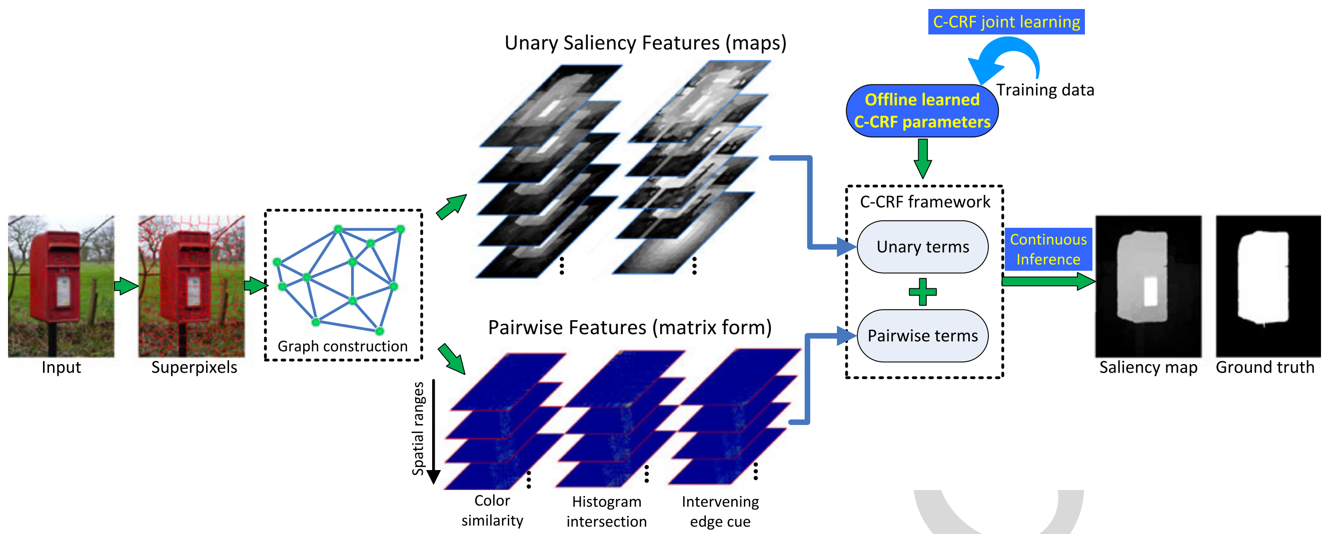

Fig. 2. Block diagram of the proposed salient object detection method.

only a small number of graph nodes are needed. An input image $\mathbf{x}$ is first over-segmented into n superpixels by using the SLIC algorithm [48], which is very widely employed by previous work [25], [27], [37], [49]–[51] as a pre-processing step. Then superpixels are used as processing units. A graph $G = (V, E)$ is then constructed, where the node set V consists of superpixels. In this paper, the terms of "superpixels" and "graph nodes" are interchangeable, and $v_i, i \in \{1 : n\}$ indicates the ith superpixel/node. To build the connections of graph edges, we first construct an initial adjacency graph $G^0 = (V, E^0)$, where vertices V correspond to superpixels. E^0 is the edge set (weighed by value 1.0) formed between pairs of spatially adjacent superpixels. Let $D^0(v_i, v_j)$ be the length of the shortest path on G^0 between nodes v_i and v_j. Then, the edge set E is formed between pairs of superpixels that are less than T nodes away on G^0, namely

$$e_{ij} \in E, \text{ if } D^0(v_i, v_j) \leq T \tag{4}$$

where T ($T \geq 1$) is a predefined integer that specifies the *maximum spatial range*. Further, as observed in many images that boundary superpixels are likely the same semantic background and also inspired by previous work [25], [50]–[52], we establish connection between arbitrary boundary superpixels as below

$$e_{ij} \in E, \text{ if } v_i, v_j \in \mathbb{B} \tag{5}$$

where $\mathbb{B}$ is a set containing all boundary superpixels. Fig. 3 shows an example of the graph connections for the case of $T = 3$. By this mean, boundary superpixels are able to serve as "bridges" for labeling consistency in image background.

B. C-CRF Composition

The definitions of unary term $U_{\boldsymbol{\alpha}}$ and pairwise term $P_{\boldsymbol{\varphi}}$ in our method are motivated by the work of Qin *et al.* [28]. The basic idea is that although a unary term calculates the dependency between a node label y_i and the entire image $\mathbf{x}$ (Fig. 1), the case can be simplified by considering the dependency between y_i and a corresponding feature vector $\mathbf{f}_i$ that derives from $\mathbf{x}$. In our case, $\mathbf{f}_i$ is a feature vector that captures the saliency information in $\mathbf{x}$ (see Section IV-C).

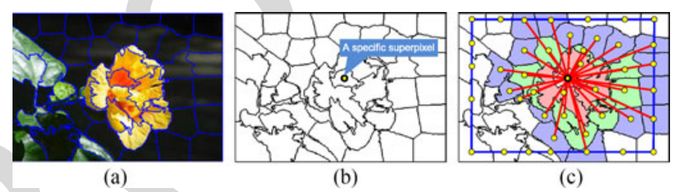

Fig. 3. Superpixels and graph construction. (a) An input image with superpixel boundaries overlapped in blue. About 50 superpixels are generated just for better illustration. (b) Superpixel boundaries in black are shown, where a superpixel is specified. (c) C-CRF graph connections, where the connections (red lines) from the specific node are shown. The maximum spatial range $T = 3$ is set as example. Superpixels filled in red/green/blue mean that they are 1/2/3 nodes away from the specific superpixel, respectively. Besides, blue lines around image boundary means that two arbitrary boundary nodes are connected, as expressed in (5).

The unary term: Assuming a d-dimensional unary saliency feature vector $\mathbf{f}_i$ for node v_i, the unary term is defined as a weighted sum of quadratic cost

$$U_{\boldsymbol{\alpha}}(y_i, \mathbf{x}) = \sum_{k=1}^{d} \alpha_k (y_i - f_{i,k})^2 \tag{6}$$

where α_k, $f_{i,k}$ are the kth components of $\boldsymbol{\alpha}$ and $\mathbf{f}_i$ respectively, and α_k indicates the weight of the kth component in the feature vector. The overall cost becomes larger if the label y_i deviates from the correspondent feature components with high weights. Further, in (6), $\mathbf{x}$ is omitted for simplicity, though the unary feature vector $\mathbf{f}_i$ is dependent on $\mathbf{x}$.

The pairwise term: Likewise, the pairwise term is a weighted sum of quadratic cost defined as

$$P_{\boldsymbol{\varphi}}(y_i, y_j, \mathbf{x}) = \frac{1}{2} \sum_{k=1}^{h} \varphi_k S_{ij}^k (y_i - y_j)^2 \tag{7}$$

where S_{ij}^k is the kth pairwise feature defined between nodes v_i and v_j, φ_k is the kth component of $\boldsymbol{\varphi}$, and h is the number of pairwise features. In the proposed method, S_{ij}^k is a positive affinity (similarity) function between v_i and v_j, and it is large if v_i and v_j are similar, so that they can be assigned with similar labels by C-CRF. Similar to (6), we have omitted $\mathbf{x}$ for simplicity in (7), although S_{ij}^k depends on $\mathbf{x}$ as well.

The energy function: According to (2), the energy function $\mathbb{E}(\mathbf{y},\mathbf{x})$ has the following form:

$$\mathbb{E}(\mathbf{y},\mathbf{x})=\sum_{i=1}^{n}\sum_{k=1}^{d}\alpha_k(y_i-f_{i,k})^2+\sum_{i,j,i\sim j}\frac{1}{2}\sum_{k=1}^{h}\varphi_k S_{ij}^k(y_i-y_j)^2 \tag{8}$$

where $\alpha_k > 0$ and $\varphi_k \geq 0$ are needed to ensure the partition function $Z(\mathbf{x})$ analytically computable (will be clear later).

Let $\mathbf{F}$ denote the stacked *feature matrix* whose row is $\mathbf{f}_i^T$, and let $\mathbf{S}^k$ denote the matrix whose entry is S_{ij}^k. With some mathematic derivation, the matrix form of (8) can be equivalently expressed as

$$\begin{aligned}\mathbb{E}(\mathbf{y},\mathbf{x}) &= \mathbf{e}^{\mathrm{T}}\boldsymbol{\alpha}\mathbf{y}^{\mathrm{T}}\mathbf{I}\mathbf{y}-2\mathbf{y}^{\mathrm{T}}\mathbf{F}\boldsymbol{\alpha}+\mathrm{Tr}\{\mathbf{F}\mathrm{diag}(\boldsymbol{\alpha})\mathbf{F}^{\mathrm{T}}\}\\ &\quad+\sum_{k=1}^{h}\varphi_k\mathbf{y}^{\mathrm{T}}\mathbf{L}^k\mathbf{y}\end{aligned} \tag{9}$$

where $\mathbf{e}$ is an all-one vector, $\mathbf{I}$ is an identity matrix, $\mathrm{Tr}\{\cdot\}$ is the trace, $\mathrm{diag}(\boldsymbol{\alpha})$ is the diagonal matrix with $\boldsymbol{\alpha}$ in the diagonal, and $\mathbf{L}^k$ is the Laplacian matrix of $\mathbf{S}^k$. The definition of Laplacian matrix is $\mathbf{L}^k := \mathbf{D}^k - \mathbf{S}^k$, where $\mathbf{D}^k$ is the degree matrix whose ith diagonal entry is $D_{ii}^k = \sum_j S_{ij}^k$.

The partition function: The partition function $Z(\mathbf{x})$ in the proposed scheme is integrable due to the continuous property of C-CRF. Firstly we introduce the below notation $\mathbf{A}, \mathbf{b}, c$:

$$\mathbf{A}=\mathbf{e}^{\mathrm{T}}\boldsymbol{\alpha}\mathbf{I}+\sum_{k=1}^{h}\varphi_k\mathbf{L}^k,\ \mathbf{b}=\mathbf{F}\boldsymbol{\alpha},\ c=\mathrm{Tr}\{\mathbf{F}\mathrm{diag}(\boldsymbol{\alpha})\mathbf{F}^{\mathrm{T}}\}. \tag{10}$$

Then according to the Gaussian integration [53], we have

$$\begin{aligned}Z(\mathbf{x}) &= \int\exp(-\mathbb{E}(\mathbf{y},\mathbf{x}))d\mathbf{y}\\ &= \exp(-c)\int\exp(-\mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{y}+2\mathbf{y}^{\mathrm{T}}\mathbf{b})d\mathbf{y}\\ &= \exp(-c)\int\exp(-\frac{1}{2}\mathbf{y}^{\mathrm{T}}(2\mathbf{A})\mathbf{y}+(2\mathbf{b})^{\mathrm{T}}\mathbf{y})d\mathbf{y}\\ &= \frac{\pi^{\frac{n}{2}}}{|\mathbf{A}|^{\frac{1}{2}}}\exp(\mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{b}-c)\end{aligned} \tag{11}$$

where n equals to the dimension of $\mathbf{A}$, and $|\mathbf{A}|$ is the determinant. The invertibility of $\mathbf{A}$ is guaranteed, as $\alpha_k > 0$, $\varphi_k \geq 0$, and $\mathbf{L}^k$ is positive semi-definite.

The negative log-likelihood: Substitute (11) and (9) into (3) meanwhile notice the notations in (10), the negative log likelihood in (3) can be re-written as

$$\begin{aligned}&-\log p(\mathbf{y}|\mathbf{x})\\ &=\mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{y}-2\mathbf{y}^{\mathrm{T}}\mathbf{b}+\mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{b}+\frac{n}{2}\log\pi-\frac{1}{2}\log|\mathbf{A}|.\end{aligned} \tag{12}$$

TABLE II
COLUMNS OF THE UNARY SALIENCY FEATURE MATRIX $\mathbf{F}$

Column	Categorization	Description
$\mathbf{F}_{:,1\sim4}$	Connectivity-based	Geodesic distance to each side of image borders
$\mathbf{F}_{:,5}$	Connectivity-based	Minimum geodesic distance to four image borders
$\mathbf{F}_{:,6}$	Connectivity-based	Normalized soft region area subtracted by 1
$\mathbf{F}_{:,7}$	Contrast-based	Spatially weighted color contrast to other superpixels
$\mathbf{F}_{:,8}$	Contrast-based	Color contrast to all boundary superpixels (backgroundness)
$\mathbf{F}_{:,9}$	Distribution heuristic	Normalized color spatial variances subtracted by 1
$\mathbf{F}_{:,10}$	Distribution heuristic	Image center bias map
$\mathbf{F}_{:,11}$	Clarity-based	Normalized singular value feature subtracted by 1

TABLE III
PAIRWISE FEATURES (IN MATRIX FORM) BETWEEN SUPERPIXELS FROM EDGE SETS E_B AND $E_{x|x\in\{1:T\}}$

Notation	Categorization	Description
$\mathbf{S}^1, \mathbf{S}^{2\sim T+1}$	Color-based	Color similarity ($S_{ij}^{(c)}$) from E_B and $E_{x|x\in\{1:T\}}$
$\mathbf{S}^{T+2}, \mathbf{S}^{T+3\sim2T+2}$	Color-based	Histogram intersection ($S_{ij}^{(h)}$) from E_B and $E_{x|x\in\{1:T\}}$
$\mathbf{S}^{2T+3}, \mathbf{S}^{2T+4\sim3T+3}$	Edge-based	Intervening edge cue ($S_{ij}^{(e)}$) from E_B and $E_{x|x\in\{1:T\}}$

C. Definition of Unary and Pairwise Features

This subsection describes the unary saliency features ($\mathbf{f}_i$) and the pairwise features (S_{ij}^k) in the proposed C-CRF model. The proposed formulation of pairwise potentials enables learning importantce for different spatial ranges of graph connections. All features used are summarized in Tables II and III. Details are given below:

1) Unary Saliency Features: Unary saliency feature vector $\mathbf{f}_i \in \mathbb{R}^d$ is initial description for the saliency level of v_i. Each component of $\mathbf{f}_i$ is a type of pre-computed saliency feature, where regions correspond to larger components are more salient. Recall that $\mathbf{F} \in \mathbb{R}^{n\times d}$ is the feature matrix whose row is $\mathbf{f}_i^{\mathrm{T}}$. Thereby a certain column of $\mathbf{F}$ can be regarded as a type of *feature map*, denoted as $\mathbf{F}_{:,k}, k \in \{1:d\}$ (Fig. 4). The unary saliency features considered in this paper fall into four types: connectivity-based, contrast-based, distribution heuristics, and clarity-based features, as given in Table II.

Connectivity-based features: Connected regions tend to be perceived as one entity by human eyes, and regions that easily connect to the image boundary are likely to be the background [49]. The boundary connectivity is hence defined based on the geodesic distance [49]. Computing geodesic distance between superpixels and four image borders separately leads to four feature maps, denoted as $\mathbf{F}_{:,1\sim4}$. The minimum geodesic distance between superpixels and image boundary leads to a single feature map $\mathbf{F}_{:,5}$. Since salient objects usually occupy small regions as comparing to large areas of background, we compute

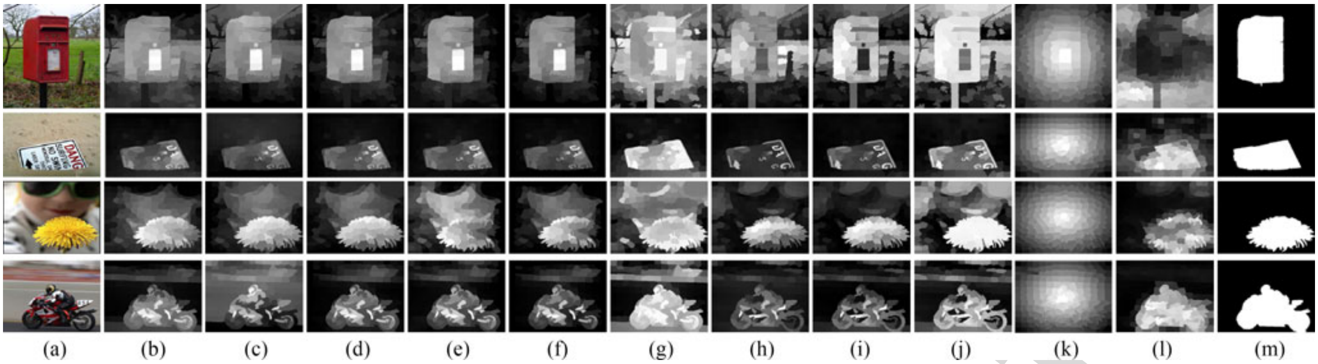

Fig. 4. Feature maps of unary saliency features. From left to right: (a) Input images. (b)–(g) Features $\mathbf{F}_{:,1\sim 6}$ (connectivity-based features). (h)–(i) Features $\mathbf{F}_{:,7\sim 8}$ (contrast-based features). (j)–(k) Features $\mathbf{F}_{:,9\sim 10}$ (distribution-based features). (l) Feature $\mathbf{F}_{:,11}$ (clarity-based feature). (m) Ground truth masks.

a type of feature which takes the region size into account. Let $d_{geo}(v_i, v_j)$ be the geodesic distance between superpixel v_i and v_j, the geodesic affinity [50] between v_i and v_j can be defined as $\mathcal{A}_{ij} = \exp(-\frac{d^2_{geo}(v_i,v_j)}{2\sigma_g^2})$ (σ_g is set according to [50]). Then we compute the spanning area associated with v_i as $\sum_{j=1}^{n} \mathcal{A}_{ij}$. This definition of the region area avoids an explicit hard segmentation of image and is "soft". Meanwhile it takes advantage of superpixels. Finally, $\mathbf{F}_{:,6}$ is formed by normalizing the spanning area value within the range [0, 1.0] and then subtracting the result from 1.0. This fits the intuition that small object regions tend to be salient (Fig. 4(g).

Contrast-based features: Global color contrast is an indicator for saliency [8], [54]. $\mathbf{F}_{:,7}$ is computed similarly to [54] by comparing the color contrast of a superpixel to other superpixels, where spatially nearer superpixels are rendered larger weights. Furthermore, we formulate $\mathbf{F}_{:,8}$ by computing the contrast of a superpixel v_i to all the boundary superpixels as $\sum_{v_j \in \mathbb{B}} ||\mathbf{c}_i - \mathbf{c}_j||_2$, where $\mathbf{c}_i$ and $\mathbf{c}_j$ are the average colors of superpixel v_i and v_j, since boundary superpixels are likely to be the background [43], [49].

Distribution heuristics: Salient objects tend to present compact color distribution [54], [55]. Taking into consideration of this, we compute a color distribution map [54], where spatial variances of colors are normalized and subtracted by 1.0 to form $\mathbf{F}_{:,9}$. Furthermore, to describe the center-bias in human attention [39], [56], $\mathbf{F}_{:,10}$ in our case is a parameter-free center-bias map computed by

$$f_{i,10} = 1 - \frac{||\mathbf{p}_i - \mathbf{p}_c||_2}{\sqrt{(l_h/2)^2 + (l_w/2)^2}} \tag{13}$$

where l_h, l_w are the height and width of the image, $\mathbf{p}_i$, $\mathbf{p}_c$ are the spatial coordinates of v_i and image center, respectively.

Clarity-based feature: Photographers tend to put objects of interest in focus meanwhile defocus irrelevant background when making high quality photos. To characterize this, we consider the Singular Value Feature (SVF) [57], [58] that models the degree of blur. An input image is first split into $l \times l$ number of grids, and the SVF [57] is then computed from each grid and further assigned to the pixels in the grid. A superpixel-based map is obtained by averaging SVF of pixels in every superpixel. After normalizing all SVF values into [0, 1.0], they are subtracted by 1 to achieve the final feature map $\mathbf{F}_{:,11}$. This describes focused objects as more salient. It is worth noting that $l \in \{10, 20, 30\}$ are used to consider different scales, and feature maps are averaged to form $\mathbf{F}_{:,11}$ [Fig. 4(l)].

Remarks: In total, unary saliency feature vector $\mathbf{f}_i$ has 11 components (i.e., $d = 11$), with each dimension normalized into [0, 1.0]. Fig. 4 shows examples of all feature maps visually. Noting that some of the features mentioned above are employed in existing work, however with different application context. We reformulate and modify the above features to constitute a unary feature matrix $\mathbf{F}$ for our C-CRF model. It is worthy noting our model is generic and not limited to the above features. If needed, more features can be easily integrated in such a way.

2) Pairwise Features: As summarized in Table III, we consider color-based and edge-based pairwise features to capture the interaction between superpixels.

Color-based features: For color-based pairwise features, we consider the average-color similarity $S^{(c)}_{ij} = \mathrm{e}^{-\lambda_c ||\mathbf{c}_i - \mathbf{c}_j||_2}$ and the histogram intersection $S^{(h)}_{ij} = \sum_{k=1} \min\{\hbar_{i,k}, \hbar_{j,k}\}$, where $\mathbf{c}_i$ and $\mathbf{c}_j$ are the average colors of v_i and v_j, and $\hbar_i, \hbar_j$ are the normalized color histograms from v_i and v_j. We obtain quantized color histograms similarly to Cheng's work [59] by first dividing the color space into $8^3 = 512$ bins. Color bins that are occupied by 99% of image pixels are kept, whereas pixels with discarded colors are then replaced by their nearest colors. This reduces the dimension of histograms and makes the computation more efficient.

Edge-based features: The edge-based feature is defined as $S^{(e)}_{ij} = \mathrm{e}^{-\lambda_e \max_{i' \in \bar{ij}} ||f_{i'}||}$, where $\bar{ij}$ is a straight line connecting centers of v_i and v_j on the image plane, i' is a pixel on $\bar{ij}$, and $||f_{i'}||$ is the edge magnitude at i' that can be derived from some edge detector. The rationale behind this feature is that $S^{(e)}_{ij}$ becomes small when there exists strong intervening edges between two superpixels, meaning v_i and v_j are less likely to have similar labels. We adopt the structured random forest-based edge detector proposed in [60] as it produces multi-scale edges with fast speed. It is worthy noting that in practice, although most superpixels in an image have their centroids inside due to the spatial compactness of SLIC superpixels [48], there may be few

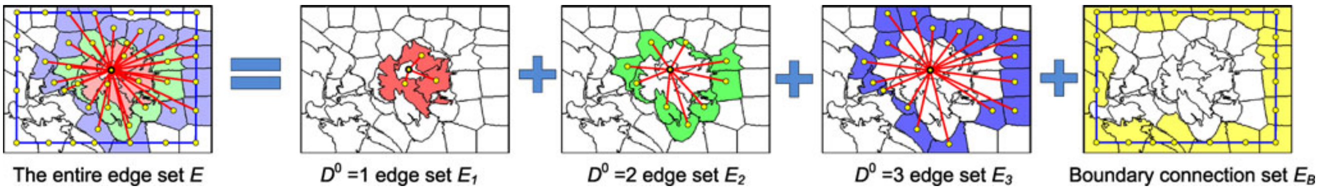

Fig. 5. Decomposition of graph edge connections in Fig. 3. The graph topology is decomposed into a boundary set E_B and $D^0 = 1, 2, 3$ sets E_1, E_2, E_3, which indicates that superpixels which are exactly one, two, and three nodes away are connected.

superpixels whose centroids are located out of them, resulting in less accurate intervening edge cue. Hence for such a case, a pixel location is sampled randomly inside each superpixel, and is used instead of the centroid to compute the edge-based features.

Graph connection decomposition: To enable learning different importance of different spatial ranges, the initial graph topology of G is partitioned into $(T+1)$ edge sets $E_B, E_1, E_2, \ldots, E_T$, where E_B contains only boundary connections, whereas $E_{x|x\in\{1:T\}}$ contains connections between superpixels that are exactly x nodes away. Such a graph decomposition is designed for properly representing different spatial ranges meanwhile avoiding an individual edge being counted multiple times. An example of such topology decomposition is shown in Fig. 5, where $T = 3$. For a specific type of connections $E_{x|x\in\{B,1,2,\ldots,T\}}$, three aforementioned pairwise features are calculated, leading to $3 \times (T+1)$ pairwise potentials (see Table III). For example, when specifying the maximum range $T = 3$ (Figs. 3 and 5), it typically results in 12 pairwise features corresponding to 12 matrices ($\mathbf{S}^{1\sim 12}$). The proposed graph decomposition enables C-CRF to automatically learn different weights for different ranges of connections. $\varphi_k \to 0$ is equivalent to discarding a type of connections if their contribution is very little during learning.

Remarks: Although some of the pairwise information above is employed by existing saliency work to build graph weights, they are usually used in an unsupervised fashion. In contrast, we combine the above features in a supervised way through learning a complete C-CRF. Besides, the advantage of our formulation of pairwise potentials is that it avoids the manual effort of tuning spatial connections. It has been observed in recent work [25], [27], [61], [62] that the ranges of spatial connections impact the final detection performance. Most of those models typically adopt non-local graph connections which are manually determined. Choosing appropriate graph connections, however, is a non-trivial task and the optimal connection ranges can depend on the coarseness of superpixels in the image. By contrast, our technique enables one to specify a relatively large maximum range T and then automatically learn the corresponding weights of connections within T. By checking the weights, one can further decide whether extension or pruning of spatial ranges is needed.

D. C-CRF Learning and Inference

We formulate the C-CRF learning as follows: given N training images $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N$ with their ground truth labels $\mathbf{y}^1, \mathbf{y}^2, \ldots, \mathbf{y}^N$, learn C-CRF parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\varphi}$. The regularized maximum conditional likelihood (RMCL) training is adopted for C-CRF learning, which is equivalent to minimizing (3) summed over all training images

$$\min_{\boldsymbol{\alpha},\boldsymbol{\varphi}} \sum_{i=1}^{N} \left\{ -\log p(\mathbf{y}^i|\mathbf{x}^i) + \frac{\lambda_1}{2}||\boldsymbol{\alpha}||_2^2 + \frac{\lambda_2}{2}||\boldsymbol{\varphi}||_2^2 \right\} \quad \text{s.t. } \alpha_k > 0,\ \varphi_k \geq 0 \tag{14}$$

where λ_1 and λ_2 are regularization parameters (pre-tuned). The optimal solution can be found by using gradient descent [28], [29]. Due to the constraints $\alpha_k > 0$ and $\varphi_k \geq 0$, we apply gradient descent iteratively on $\log \alpha_k$ and $\log \varphi_k$ during the optimization. Let the gradient of the energy loss in (14) w.r.t. $\log \alpha_k$ and $\log \varphi_k$ be $\nabla_{\log \alpha_k}$ and $\nabla_{\log \varphi_k}$, respectively. Here by dropping the summation operation for notation simplicity, the derivation of $\nabla_{\log \alpha_k}$ and $\nabla_{\log \varphi_k}$ is written as

$$\nabla_{\log \alpha_k} = \alpha_k \left\{ \sum_i (y_i - f_{i,k})^2 + \frac{\partial \log Z(\mathbf{x})}{\partial \alpha_k} + \lambda_1 \alpha_k \right\} \tag{15}$$

$$\nabla_{\log \varphi_k} = \varphi_k \left\{ \mathbf{y}^{\mathrm{T}} \mathbf{L}^k \mathbf{y} + \frac{\partial \log Z(\mathbf{x})}{\partial \varphi_k} + \lambda_2 \varphi_k \right\} \tag{16}$$

where further according to (11) and use the notations in (10), $\frac{\partial \log Z(\mathbf{x})}{\partial \alpha_k}$ can be computed

$$\frac{\partial \log Z(\mathbf{x})}{\partial \alpha_k} = -\frac{1}{2|\mathbf{A}|}\frac{\partial |\mathbf{A}|}{\partial \alpha_k} + \frac{\partial \mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{b}}{\partial \alpha_k} - \frac{\partial c}{\partial \alpha_k} \tag{17}$$

$$\frac{\partial |\mathbf{A}|}{\partial \alpha_k} = |\mathbf{A}|\mathrm{Tr}(\mathbf{A}^{-1}) \tag{18}$$

$$\frac{\partial \mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{b}}{\partial \alpha_k} = \mathbf{F}_{:,k}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{b} - \mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{A}^{-1}\mathbf{b} + \mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{F}_{:,k} \tag{19}$$

$$\frac{\partial c}{\partial \alpha_k} = ||\mathbf{F}_{:,k}||_2^2 \tag{20}$$

and $\frac{\partial \log Z(\mathbf{x})}{\partial \varphi_k}$ can be computed

$$\frac{\partial \log Z(\mathbf{x})}{\partial \varphi_k} = -\frac{1}{2|\mathbf{A}|}\frac{\partial |\mathbf{A}|}{\partial \varphi_k} + \frac{\partial \mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{b}}{\partial \varphi_k} \tag{21}$$

$$\frac{\partial |\mathbf{A}|}{\partial \varphi_k} = |\mathbf{A}|\mathrm{Tr}(\mathbf{A}^{-1}\mathbf{L}^k) \tag{22}$$

$$\frac{\partial \mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{b}}{\partial \varphi_k} = -\mathbf{b}^{\mathrm{T}}\mathbf{A}^{-1}\mathbf{L}^k\mathbf{A}^{-1}\mathbf{b}. \tag{23}$$

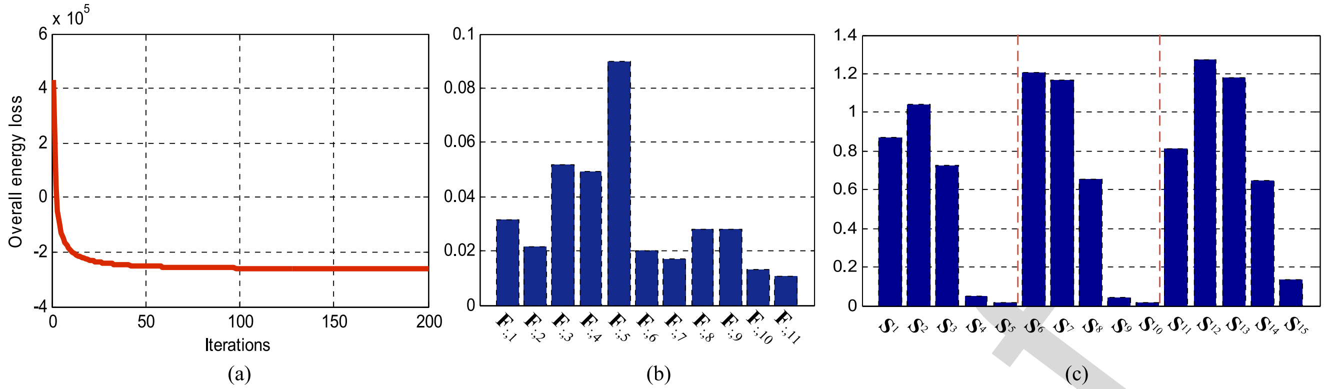

Fig. 6. C-CRF learning outcomes. (a) The overall energy changes of (14) regarding to iterations. (b) The learned $\boldsymbol{\alpha}$. (c) The learned $\boldsymbol{\varphi}$ ($T = 4$ case). The notations of features in (b) and (c) are consistent with those in Tables II and III.

When $\boldsymbol{\alpha}$ and $\boldsymbol{\varphi}$ are learned, the saliency inference of C-CRF on a new test image is achieved by minimizing (9). Setting $\partial \mathbb{E}(\mathbf{y}, \mathbf{x})/\partial \mathbf{y} = 0$ leads to the closed-form solution

$$\begin{aligned} \mathbf{y} &= \left(\mathbf{e}^{\mathrm{T}} \boldsymbol{\alpha} \mathbf{I} + \sum_{k=1}^{h} \varphi_k \mathbf{L}^k \right)^{-1} \mathbf{F} \boldsymbol{\alpha} \\ &= \mathbf{A}^{-1} \mathbf{b}. \end{aligned} \tag{24}$$

In (24), the invertibility is guaranteed, since $\alpha_k > 0$, $\varphi_k \geq 0$, and $\mathbf{L}^k$ is positive semi-definite. Finally, we normalize $\mathbf{y}$ into [0, 1.0] to render a final saliency map. After the normalization, we ensure at least one superpixel has value 1 and at least 10% superpixels have values 0.

A diffusion perspective of C-CRF: Interestingly, the above closed-form solution of the C-CRF inference coincides with the unified formulation of diffusion-based saliency methods raised in [27], where $(\mathbf{e}^{\mathrm{T}} \boldsymbol{\alpha} \mathbf{I} + \sum_{k=1}^{h} \varphi_k \mathbf{L}^k)^{-1} = \mathbf{A}^{-1}$ is the diffusion matrix and $\mathbf{F}\boldsymbol{\alpha}$ is the integrated saliency "seed vector" that leads to a *raw saliency map*. The optimal solution (inference) is the product of the diffusion matrix and a saliency seed vector, leading to the equilibrium state vector as the diffused saliency detection results. However, comparing to [27], we have investigated a different framework—C-CRF, with a totally different learning strategy.

V. Experiments and Results

A. Setup

Six benchmark datasets were used for our tests, including: ASD [7] (1000 images), MSRA-B [24] (5000 images), ECSSD [56] (1000 images), SOD [63] (300 images), SED1 (one-object image set having 100 images) [64], and SED2 (two-objects image set having 100 images). Training images were chosen from MSRA-B. Since the ASD dataset is a subset of MSRA-B, in order to evaluate the performance on ASD, we first exclude images that belong to ASD from MSRA-B, resulting in 4000 images remained. Then we randomly select 3000 images for training and leave the other 1000 images as the MSRA-B test set. The trained C-CRF on this dataset is then applied to other datasets. C-CRF parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\varphi}$ to be learned were initialized as all-one vectors. The regularization parameters $\lambda_1 = 1$ and $\lambda_2 = 5$ were set. Since performing the gradient descent on 3000 training samples are tractable, we used the gradient descent instead of stochastic gradient descent for learning, in order to achieve more stable convergence. The learning rate was set as 1×10^{-5}, and the convergence was achieved after 200 iterations.

During feature extraction, each image was segmented into $n \approx 200$ superpixels. The maximum graph range T was initially set to 4 but then pruned to 3 according to the learning outcomes (see Section V-B for details). Besides, all parameters for individual unary and pairwise features were empirically set ($\lambda_e = 10$ and $\lambda_c = 10$ were set for pairwise features).

B. Learning Outcomes

Fig. 6 shows the learning results of the C-CRF. From Fig. 6(a), one can see that the overall energy decreases monotonously as gradient descent proceeds and has reached a stable minimum. Due to the continuous property of C-CRF, (1) computed on some images might be larger than 1.0 and it would result in a negative log-likelihood. This is why in a) the overall energy turns negative as iteration proceeds. This phenomenon on C-CRF is different from D-CRF since the solution space of the latter is finite and countable. Hence for D-CRF, (1) will result in a probability value instead of a probability density value.

Fig. 6(b) and 6(c) show the learned $\boldsymbol{\alpha}$ and $\boldsymbol{\varphi}$, respectively. The learning results in Fig. 6(b) indicate that the geodesic features $\mathbf{F}_{:,1\sim5}$ are the most informative ones, which have gained large weights. Among them $\mathbf{F}_{:,5}$ is the most important one. Following that, the contrast to image boundary ($\mathbf{F}_{:,8}$) and color spatial distribution ($\mathbf{F}_{:,9}$) gain larger weight than the global contrast ($\mathbf{F}_{:,7}$) and center-bias ($\mathbf{F}_{:,10}$). This observation is somewhat consistent with [27] where the center bias feature does not appear in the top among the listed features. The last feature $\mathbf{F}_{:,11}$ (clarity-based) has obtained the lowest weight. The cause of this is that like most blur detectors, the SVF is based on local gradient and has limitation in distinguishing between smooth object surfaces and real blurred image regions [e.g., the 1st row of Fig. 4(l)].

Fig. 6(c) shows the learned $\boldsymbol{\varphi}$ when setting maximum spatial range $T = 4$. One can see the learned weights of pairwise

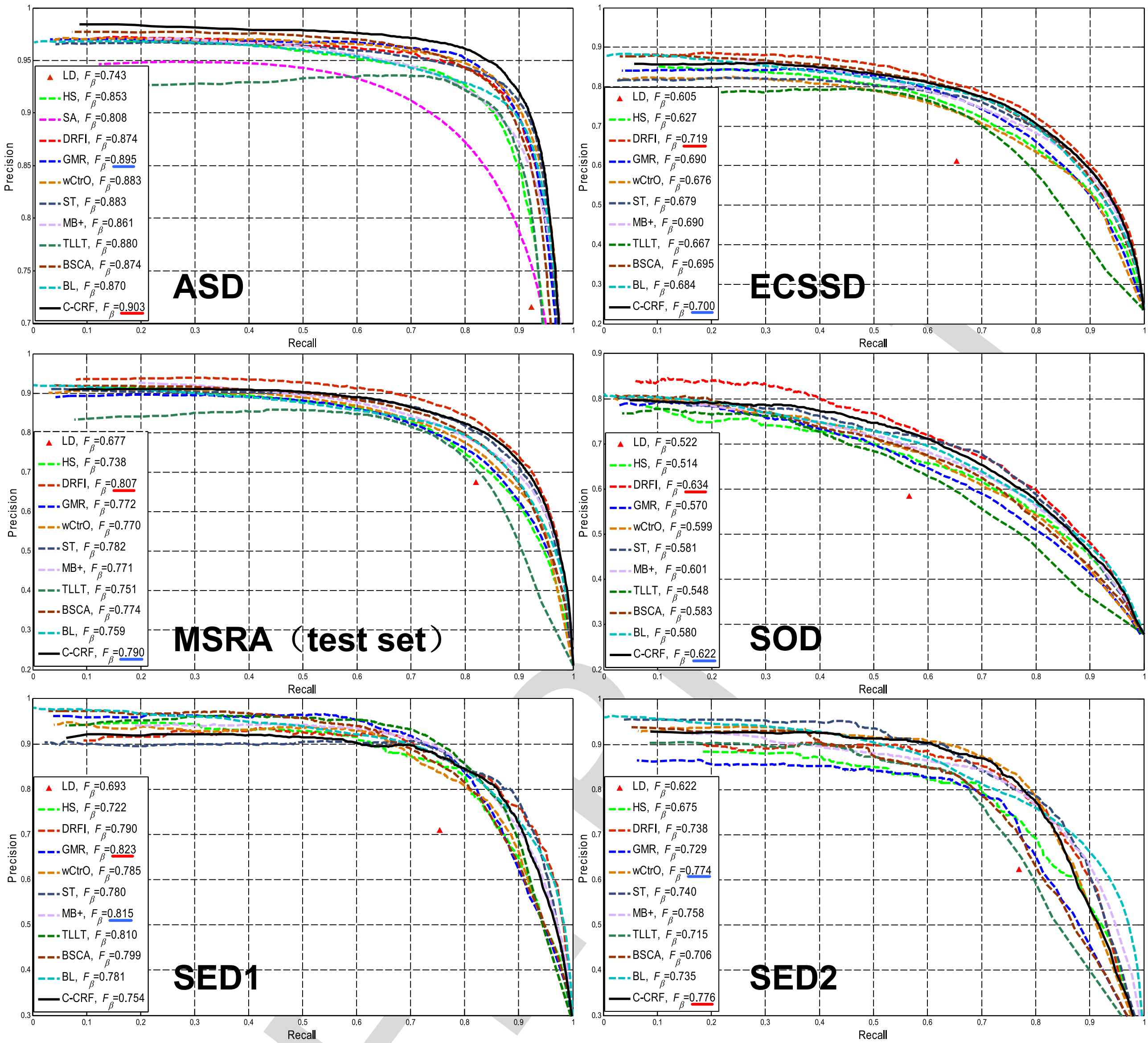

Fig. 7. Quantitative comparisons (precision-recall curves and F_β scores) of the proposed method (C-CRF) to the state-of-the-art methods on six benchmark datasets. The best and the second best F_β are underlined by red and blue.

features decrease as the spatial range D^0 increases. This meets the common sense since spatially close superpixels should have strong interaction, but noting that such relationship in our method is automatically learned rather than handcrafted. Besides, highly degraded weights (close to zeros) for features $\mathbf{S}^4$, $\mathbf{S}^9$, $\mathbf{S}^{14}$ that correspond to $D^0 = 3$ and for features $\mathbf{S}^5$, $\mathbf{S}^{10}$, $\mathbf{S}^{15}$ that correspond to $D^0 = 4$ reveal further extending the maximum spatial range to $T \geq 4$ under the current experimental setup is not essential. Considering when $D^0 = 4$, the corresponding weights are very low. In practice we prune the spatial range and use $T = 3$, as shown in Figs. 3 and 5. As shown in Fig. 6(c), the intervening edge cues ($\mathbf{S}^{11} \sim \mathbf{S}^{15}$) are the most informative ones among all pairwise features. They generally gain larger weights than color similarity ($\mathbf{S}^1 \sim \mathbf{S}^5$) and histogram intersection ($\mathbf{S}^6 \sim \mathbf{S}^{10}$). This validates that incorporating the edge cues makes contribution. Finally, large weights of boundary connections ($\mathbf{S}^1$, $\mathbf{S}^6$, $\mathbf{S}^{11}$) reveal connecting boundary superpixels is useful.

C. Comparison to Existing Methods

We compare the proposed saliency detection to 11 existing methods including: LD (Learning to Detect) [24], HS (Hierarchical Saliency) [56], SA (Saliency Aggregation) [26], DRFI (Discriminative Regional Feature Integration) [43], GMR (Graph-based Manifold Ranking) [25], wCtrO (background weighted Contrast with Optimization) [50], ST (Saliency Tree) [65], MB+ (Minimum Barrier Saliency) [66], TLLT (Teaching-to-Learn and Learning-to-Teach saliency) [51], BSCA (Background-based Single-layer Cellular Automata) [52], BL (Bootstrap Learning) [44]. Among them, LD [24], SA [26], GMR [25] are CRF-related methods listed in

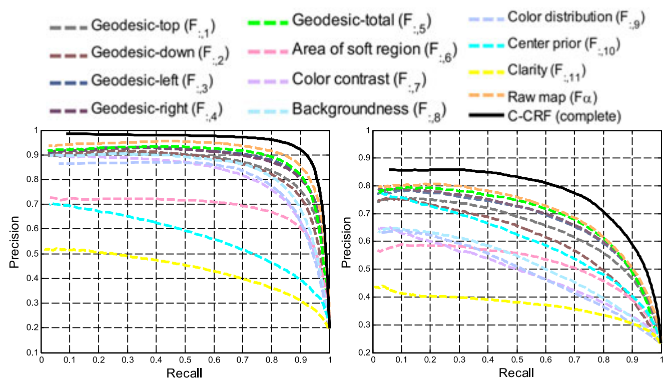

Fig. 8. Precision-recall curves of individual unary features on ASD (left) and ECSSD (right).

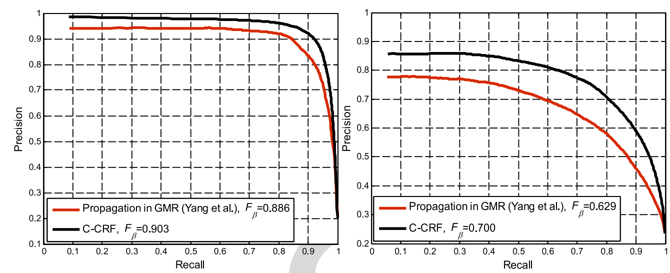

Fig. 9. Effectiveness of integrating pairwise features, validated on ASD (left) and ECSSD (right). In this test, the same "seed" vector $\mathbf{F}\boldsymbol{\alpha}$ is used.

Table I. Unfortunately, the authors of SA only provide their results on ASD dataset. Therefore, we can only evaluate SA on ASD. Besides, the code of [27] is not publicly released, so the results cannot be compared. For all the compared methods, we use the public available implementations/results provided by the authors. Precision-recall curve and F_β-measure are used for evaluating the overall performance [7], [25].

Fig. 7 shows the results of precision-recall curves and F_β scores. The proposed method (C-CRF) is comparable to state-of-the-art methods on both criteria, which has validated the effectiveness of learning a C-CRF for saliency detection. Notably, our method outperforms C-CRF related methods LD, SA, GMR together with other state-of-the-art methods with noticeable margins. Regarding to the F_β, our method consistently achieves 1st on ASD, SED2, and the 2nd on ECSSD, MSRA-B (test set) and SOD. Another data-driven method DRFI sometimes performs better than our method, which may be due to different feature extraction and learning strategies. Visual comparisons are shown in Fig. 11.

D. Integration of Features/State-of-the-Art Models

Since C-CRF is employed in this study as a principled feature integrating framework, its performance on integrating various unary and pairwise features should be evaluated. Fig. 8 shows the precision-recall curves of unary features on ASD and ECSSD. One can see that the individual features vary widely on performance, and among them $\mathbf{F}_{:,5}$ (which computes the minimum geodesic distance to image borders) achieves the best results. This coincides with the learning outcomes from MSRA-B, where $\mathbf{F}_{:,5}$ gains the highest weight. Observing Fig. 8, the weighted sum of features (the raw map computed by $\mathbf{F}\boldsymbol{\alpha}$) outperforms all individual features, but the improvement is relatively marginal. In contrast, the performance is boosted drastically by a complete C-CRF.

To validate the effectiveness of learning for pairwise features, we treat the C-CRF inference stage (24) as a diffusion process and replace its diffusion matrix $\mathbf{A}^{-1}$ with the propagation matrix used in GMR [25]. Note GMR is related to C-CRF but without learning (Table I). Its propagation matrix merely considers the similarity of average colors between superpixels, which intuitively is less effective on representing more sophisticated interaction between neighboring superpixels, such as fine-grained color information and texture differences. Fig. 9 shows the results of this experiment, which validate such an intuition. It can be seen that by using the same "seed vector" ($\mathbf{F}\boldsymbol{\alpha}$), the diffusion technique employed by [25] is inferior to C-CRF exploited in this paper.

Besides, we validate the power of integrating state-of-the-art methods by C-CRF, where 5 models are considered: HS, DRFI, GMR, wCtrO, and MB+. The resulting saliency maps from these five models are used as the unary feature maps, which are converted into superpixel-wise maps by averaging pixel-wise saliency. The C-CRF is then re-trained. Fig. 10 shows the C-CRF integration performance on ASD, MSRA (test set) and ECSSD, where the performance boost over individual methods can be observed on all three datasets. Some visual results from this experiment are in Fig. 11.

E. Effectiveness of Graph Topology Decomposition

To show the advantages of learning weights adaptively for different spatial ranges, we compare to the C-CRF variants without graph topology decomposition but with manually specified graph ranges. Here 1-ring graph, 2-ring graph, and 3-ring graph are considered. Noting an x-ring graph means a superpixel (graph node) is connected to superpixels within its x-ring neighborhood [25], [27], [61]. Besides, in each graph, the boundary superpixels are connected with each other as in this paper. For each one of the three graphs, the three types of pairwise features namely two color-based ($S_{ij}^{(c)}, S_{ij}^{(h)}$) and one image edge-based ($S_{ij}^{(e)}$) as described in Section IV-C are calculated, resulting in 3 matrices ($\mathbf{S}^{(c)}, \mathbf{S}^{(h)}, \mathbf{S}^{(e)}$) for each graph. Except for the graph ranges, all other C-CRF configurations including unary features and parameters are kept consistent with Section V-A. Then for each graph, C-CRF is re-trained and used for saliency prediction. Fig. 12 shows the quantitative comparison between the above three graphs and our graph topology decomposition. It can be observed that the proposed strategy performs more robustly than an x-ring graph which is manually specified.

F. Robustness to The Number of Superpixels

Experiments were done by varying superpixel number from 100 to 300, and meanwhile keeping other setup the same as

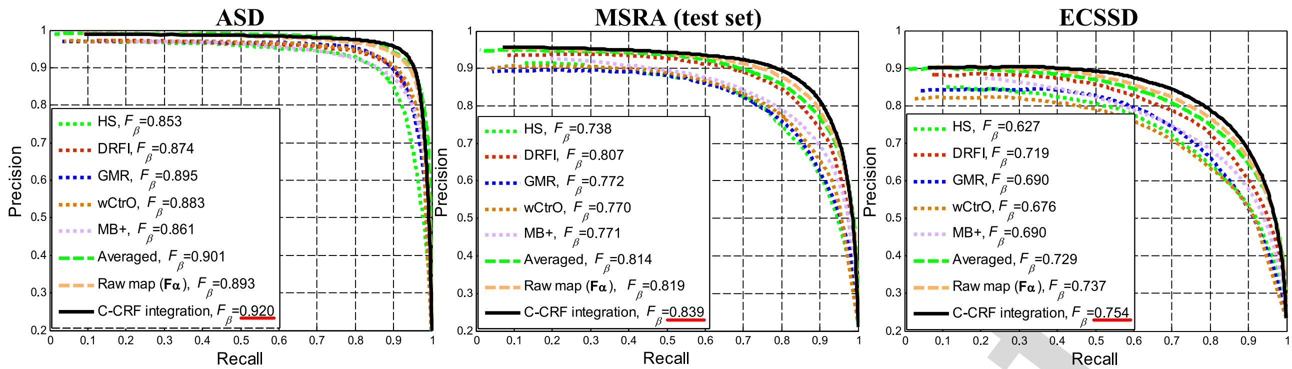

Fig. 10. Integrating five state-of-the-art methods including HS, DRFI, GMR, wCtrO, and MB+ by the proposed C-CRF based framework. The best F_β are underlined by red.

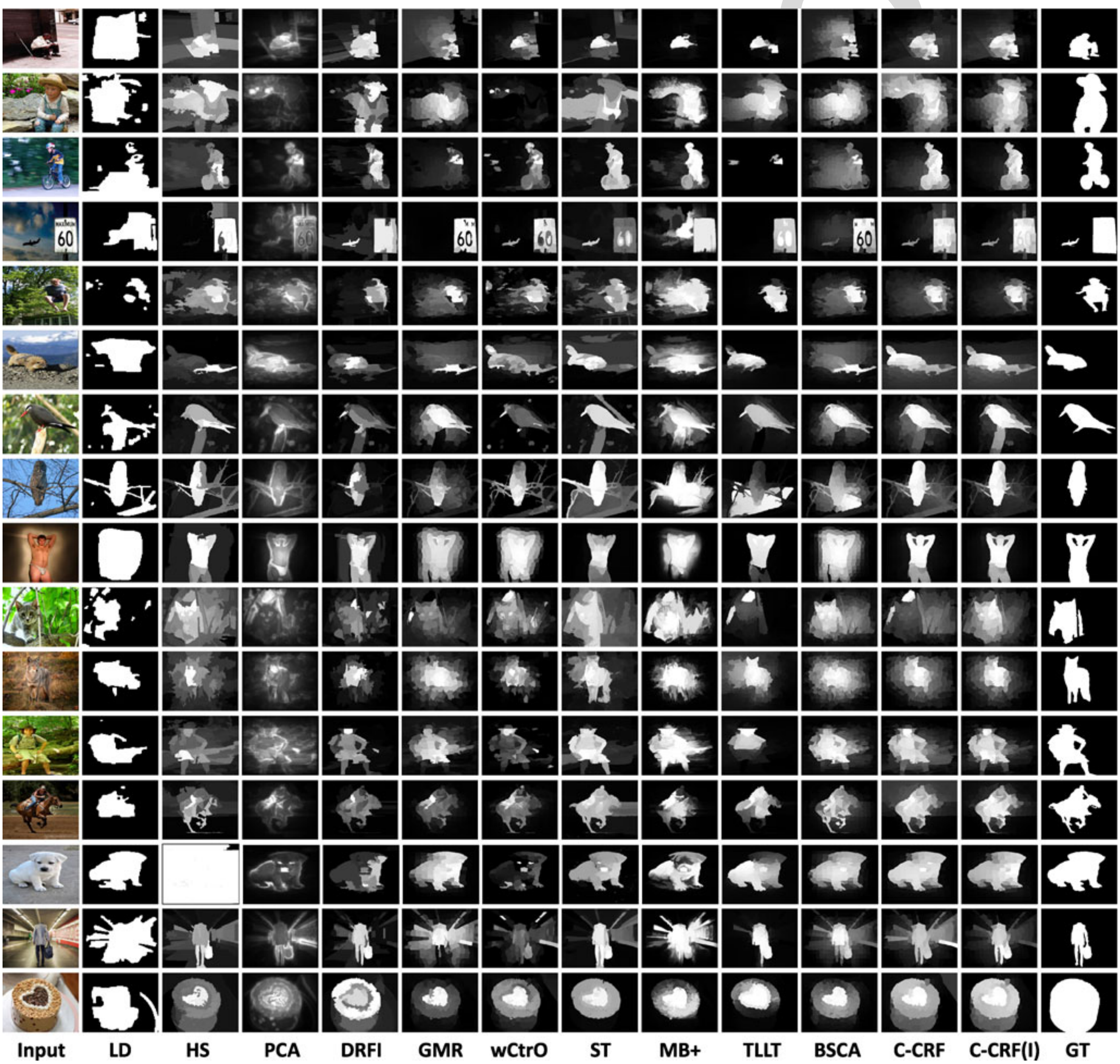

Fig. 11. Qualitative comparison of the proposed method with existing methods on some challenging images with textured background. C-CRF refers to results by integrating different unary saliency features. C-CRF(I) refers to results obtained by integrating five prior models (HS, DRFI, GMR, wCtrO, and MB+), as demonstrated in Section V-D.

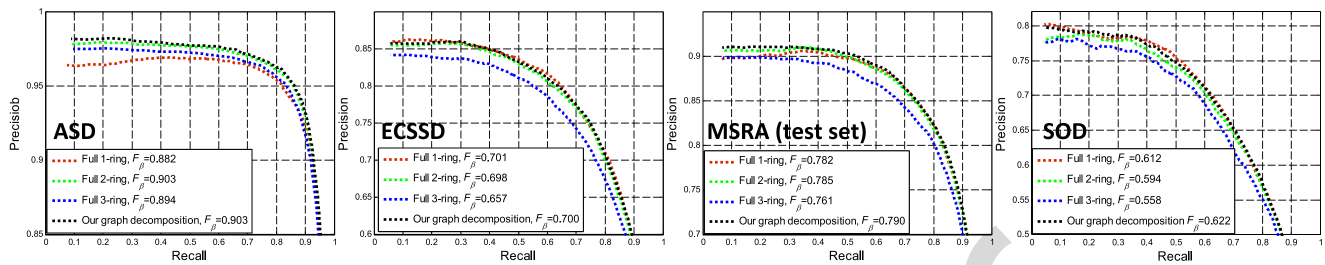

Fig. 12. Quantitative comparisons of one-ring, two-ring, and three-ring graphs and our graph topology decomposition on four benchmark datasets.

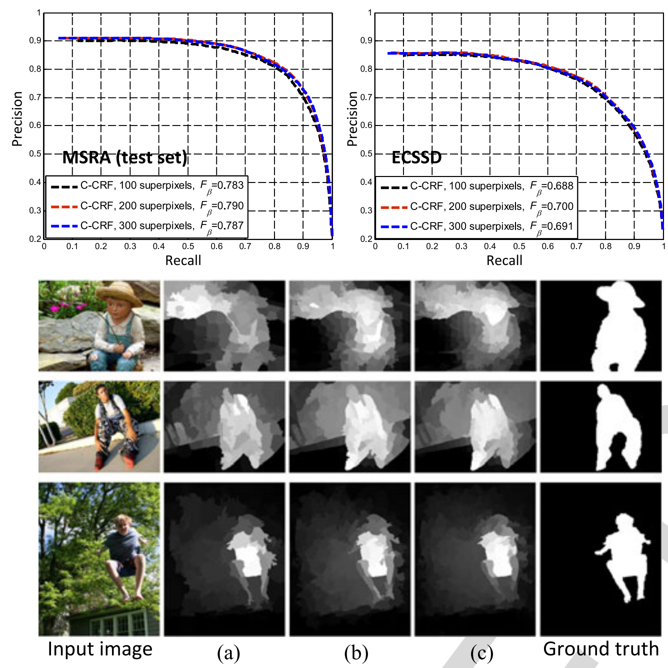

Fig. 13. Quantitative evaluation on MSRA test set (top left) and ECSSD (top right) by using different superpixel numbers. In the bottom some visual comparisons are shown: (a) 100 superpixel case, (b) 200 superpixel case, and (c) 300 superpixel case.

those mentioned in Section V-A. Next, we re-trained the C-CRF model on the training set and then tested it on MSRA-test and ECSSD. Note there are two sides of effect when varying superpixel number: 1) It somewhat affects the computed unary and pairwise features. Fewer superpixels lead to coarser image representation. Fortunately, we observed some robustness of computing unary features $\mathbf{F}_{:,1\sim 11}$ to such a change. 2) It also affects the "scale" of the C-CRF objective function, because the dimension of all vectors and matrices involved will change accordingly.

Observing the learning outcomes, we find the overall distribution (or tendency) of learned α and φ is still similar to that in Fig. 6. Fig. 13 shows the evaluation on MSRA-test and ECSSD by using different numbers of superpixels, where robustness to such change can be observed. Using 100 superpixels leads to slightly worse performance as the superpixels become coarser and hence the pre-segmentation is less accurate. Using 200 superpixels and 300 superpixels almost leads to identical performance. Some visual comparisons are shown in Fig. 13. In all, the C-CRF learning and inference is somewhat robust to superpixel number, therefore graph node numbers. No matter what setup is adopted, C-CRF will learn the optimal feature combination under the current setup.

G. Efficiency

Though the training based on gradient descent from the offline extracted features on 3000 images from MSRA-B took about 4 h, the C-CRF prediction was very fast due to the closed-form solution. It only took 2s in average to process an image from ASD dataset. The superpixel segmentation and attribute extraction (e.g., superpixel colors and histograms) took 0.4 s. The unary feature extraction took 0.45 s, and the pairwise feature extraction took 1.1s including edge detection. The running time was reported on an i7-4720HQ 2.6 GHz laptop with 8 GB memory by Matlab code without optimization.

H. Discussion About the Limitation

Though our C-CRF learning-based method enables effective feature integration and meanwhile boosts the performance from individual saliency features (Fig. 8), the major limitation is its final detection somewhat relies on the quality of input features. If none of the unary saliency features provide reasonable initial saliency estimation, the C-CRF inference will still be bad. Conversely, good features will improve the final detection. This phenomenon can be observed by comparing the quantitative results in Figs. 7 and 10, where employing the state-of-the-art results as unary features leads to better C-CRF inference. A visual example can be found in the 10th row of Fig. 11. One potential solution to this is to enrich features in the feature pool and let the C-CRF discover useful, effective ones through learning.

VI. Conclusion

This paper applies the complete learning and inference theories of continuous conditional random field (C-CRF) to salient object detection. The regularized maximum conditional likelihood training by gradient descent optimization is used for parameter learning, and the inference is achieved by an efficient closed-form solution. The power of the proposed method on integrating various unary and pairwise features is tested and evaluated comprehensively. In addition, we propose a novel formulation of pairwise features by graph topology decomposition. The effectiveness on enabling learning weights of different spatial ranges is validated with reasonable learning outcomes. Experimental results and comparison with 11 existing methods show that the proposed method achieves state-of-the-art performance on precision-recall curves with comparable F_β-measure scores. Since the proposed method enables principled feature integration, in the future some high-level features such as the category-dependent or semantic features may be incorporated into the proposed method as top-down influences.

REFERENCES

[1] A. M. Triesman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[2] C. Koch *et al.*, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence*. New York, NY, USA: Springer-Verlag, 1987, pp. 115–141.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[4] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2007, pp. 1–8.

[5] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 155–162.

[6] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

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

[8] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 409–416.

[9] Z. Liu, O. Le Meur, S. Luo, and L. Shen, "Saliency detection using regional histograms," *Opt. Lett.*, vol. 38, no. 5, pp. 700–702, 2013.

[10] A. Borji, M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[11] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 366–379.

[12] L. Wang, J. Xue, N. Zheng, and G. Hua, "Automatic salient object extraction with contextual cue," in *Proc. IEEE Int. Conf. Comput. Vis.*, Nov. 2011, pp. 105–112.

[13] Z. Liu *et al.*, "Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1275–1289, Aug. 2012.

[14] F. Stentiford, "Attention based auto image cropping," in *Proc. ICVS Workshop Comput. Attention Appl.*, 2007, pp. 1–9.

[15] L. Marchesotti *et al.*, "A framework for visual saliency detection with applications to image thumbnailing," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2232–2239.

[16] Y. Ding, X. Jing, and J. Yu, "Importance filtering for image retargeting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 89–96.

[17] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[18] S. Lin, I. Yeh, C. Lin, and T. Lee, "Patch-based image warping for content-aware retargeting," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 359–368, Feb. 2013.

[19] T. Chen, M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: Internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, 2009, Art. no. 124.

[20] Y. Gao *et al.*, "Database saliency for fast image retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 359–369, Mar. 2015.

[21] Y. Zhang, X. Qian, X. Tan, J. Han, and Y. Tang, "Sketch-based image retrieval by salient contour reinforcement," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1604–1615, Aug. 2016.

[22] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, Sep. 2014.

[23] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 88–92, Jan. 2014.

[24] T. Liu, Z. Yuan, J. Sun, J. Wang, and N. Zheng, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

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

[26] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 1131–1138.

[27] S. Lu, V. Mahadevan, and N. Vasconcelos, "Learning optimal seeds for diffusion-based salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2790–2797.

[28] T. Qin, T. Liu, X. Zhang, D. Wang, and H. Li, "Global ranking using continuous conditional random fields," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1281–1288.

[29] T. Baltrušaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using continuous conditional random fields," in *Proc. IEEE Int. Conf. Workshops Automat. Face Gesture Recog.*, Apr. 2013, pp. 1–8.

[30] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 5162–5170.

[31] A. Kolesnikov, M. Guillaumin, V. Ferrari, and C. Lampert, "Closed-form approximate CRF training for scalable image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 550–565.

[32] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vis.*, vol. 81, no. 1, pp. 2–23, 2009.

[33] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.

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

[35] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[36] S. Nowozin and C. Lampert, "Structured learning and prediction in computer vision," *Found. Trends Comput. Graph. Vis.*, vol. 6, no. 3/4, pp. 185–365, 2011.

[37] C. Yang, L. Zhang, and H. Lu, "Graph-regularized saliency detection with convex-hull-based center prior," *Signal Process. Lett.*, vol. 20, no. 7, pp. 637–640, 2013.

[38] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A survey," *CoRR*, 2014. [Online]. Available: http://arxiv.org/abs/1411.5878

[39] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2106–2113.

[40] B. Alexe *et al.*, "What is an object?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2010, pp. 73–80.

[41] P. Khuwuthyakorn, A. Robles-Kelly, and J. Zhou, "Object of interest detection by saliency learning," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 636–649.

[42] P. Mehrani and O. Veksler, "Saliency segmentation based on learning and graph cut refinement," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 1–12.

[43] H. Jiang *et al.*, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2083–2090.

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

[45] X. Zhou, Z. Liu, G. Sun, L. Ye, and X. Wang, "Improving saliency detection via multiple kernel boosting and adaptive fusion," *IEEE Signal Process. Lett.*, vol. 23, no. 4, pp. 517–521, Apr. 2016.

[46] X. Wang, L. Zhang, L. Lin, Z. Liang, and W. Zuo, "Deep joint task learning for generic object extraction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 523–531.

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

[48] R. Achanta *et al.*, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[49] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 29–42.

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

[51] C. Gong *et al.*, "Saliency propagation from simple to difficult," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 2531–2539.

[52] Y. Qin *et al.*, "Saliency detection via cellular automata," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 110–119.

[53] K. Petersen and M. Pedersen, *The Matrix Cookbook*. Lund, Sweden: Tech. Univ. Denmark, 2008.

[54] F. Perazzi, P. Krahenbul, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 733–740.

[55] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Salient region detection by modeling distributions of color and orientation," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 892–905, Aug. 2009.

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

[57] B. Su, S. Lu, and C. L. Tan, "Blurred image region detection and classification," in *Proc. ACM Conf. Multimedia*, 2011, pp. 1397–1400.

[58] J. Kim, D. Han, Y. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 883–890.

[59] M. Cheng *et al.*, "Efficient salient region detection with soft image abstraction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1529–1536.

[60] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1841–1848.

[61] J. Sun, H. Lu, and X. Liu, "Saliency region detection based on Markov absorption probabilities," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1639–1649, May 2015.

[62] K. Fu, C. Gong, I. Gu, and J. Yang, "Normalized cut-based saliency detection by adaptive multi-level region merging," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5671–5683, Dec. 2015.

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

[64] S. Alpert, M. Galun, A. Brandt, and R. Basri, "Image segmentation by probabilistic bottom-up aggregation and cue integration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 315–327, Feb. 2012.

[65] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2014.

[66] J. Zhang *et al.*, "Minimum barrier salient object detection at 80 fps," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1404–1412.

Keren Fu received the B.S. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2011, and dual Ph.D. degrees from Shanghai Jiao Tong University, Shanghai, China, and Chalmers University of Technology, Gothenburg, Sweden, in 2016, under the joint supervision of Prof. J. Yang and Prof. I. Y.-H. Gu.

His current research interests include visual computing, saliency analysis, and machine learning.

Dr. Fu was the recipient of the National Scholarship by the Ministry of Education in 2013 and 2015.

Irene Yu-Hua Gu (M'94–SM'03) received the Ph.D. degree in electrical engineering from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 1992.

From 1992 to 1996, she was a Research Fellow with the Philips Research Institute IPO, Eindhoven. The Netherlands; a Postdoc with Staffordshire University, Staffordshire, U.K.; and a Lecturer with the University of Birmingham, Birmingham, U.K. Since 1996, she has been with the Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden, where she has been a full Professor since 2004. Her research interests include statistical image and video processing, object tracking and video surveillance, pattern classification, and signal processing with applications to electric power systems.

Prof. Gu was an Associate Editor of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART A: SYSTEMS AND HUMANS and the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART B: CYBERNETICS from 2000 to 2005, and an Associate Editor of the *EURASIP Journal on Advances in Signal Processing* from 2005 to 2016. She was the Chair of the IEEE Swedish Signal Processing Chapter from 2001 to 2004. She has been on the Editorial Board of the *Journal of Ambient Intelligence and Smart Environments* since 2011.

Jie Yang received the Ph.D. degree from the Department of Computer Science, Hamburg University, Hamburg, Germany, in 1994.

He is currently a Professor with the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. He has led many research projects (e.g., National Science Foundation, 863 National High Technical Plan), published one book in Germany, and authored more than 200 journal papers. His major research interests include object detection and recognition, data fusion and data mining, and medical image processing.