This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2963701, IEEE Internet of Things Journal

Trustful Internet of Surveillance Things Based on Deeply-Represented Visual Co-Saliency Detection

Zhifan Gao, *Member, IEEE,* Chenchu Xu, Heye Zhang, *Member, IEEE,* Shuo Li, and Victor Hugo C. de Albuquerque, *Senior Member, IEEE*

***Abstract*—Trustful Internet-of-Things (IoT) plays an important role in smart cities. The trust information in surveillance data motivate the analysis of images from numerous IoT devices. Saliency detection is a fundamental step in surveillance data analysis for providing helps to the subsequent tasks, but unsuitable to IoT applications owing to the neglect of image similarity and difference from diverse IoT devices. To solve this problem, we enable the co-saliency detection in IoT, which detects the common and salient foreground regions in the group surveillance images, The main contributions include: 1) Enable a multi-stage context perception scheme to efficiently extract the contextual information corresponding to different-size receptive fields in the single image; 2) Construct a two-path information propagation to extract the inter-image similarity and difference from the high-level image feature representations of the group images; 3) Propose the stage-wise refinement to allocate the label information to different parts of the network for helping the network to learn the enriched semantically common knowledge. The extensive experiments performed on three public datasets can demonstrate the effectiveness of our approach and its superiority to four state-of-the-art co-saliency detection methods.**

I. Introduction

Building trustful Internet-of-Things (IoT) is of importance in the construction of smart cities [1], [2], [3]. It can prevent the data collection and communication process from unauthorized users and attacks of the misbehaving devices [4]. From the perspective of data trust, IoT devices can benefit from observing the data they collect, because the trust degree is closely related to the value of the information extracted by the IoT environment [5]. Specific in surveillance applications, the video/image data may imply the trust information because their visual information indicate the state of the corresponding IoT devices (e.g. expected scenes). Thus, the analysis of surveillance data may help to build the trustful IoT environment in smart cities.

Benefit from the development of artificial intelligent, the surveillance scene analysis can extract the desired information from numerous surveillance data for reducing human error, as well as lightening the burden on users. For example, the crowd anomaly detection system can automatically issue an alert when the accident occurs, without requiring the tedious manual monitoring [6], [7]. Moreover, the surveillance scene analysis from multiple IoT sources enables the event prediction and big data mining [8], [9], [10]. For instance, the surveillance images can be collected by IoT devices on vehicles and then analyzed in the cloud to enable the air quality monitoring in urban areas [11].

As a fundamental step of surveillance scene analysis, the saliency detection can help many subsequent tasks like scene cognition and understanding [12]. This is because the saliency detection is usually based on the information of the single image, which do not consider the similarity and difference of surveillance images collected by plenty of IoT devices installed at different places. This phenomenon may cause that the saliency detection has limited performance on varieties of surveillance data. It is moreover unable to provide effective help to other tasks of surveillance scene analysis.

Recent co-saliency detection is more suitable to the IoT-based surveillance scene analysis than saliency detection, because it investigates the synergetic association in many images. The co-saliency detection focuses on the detection of image foreground regions with the characteristics of similarity and saliency in a group of images [13]. It is inspired by the phenomenon that human visual attention can be attracted by the same object in the varied but similar image backgrounds [14]. Accordingly, the co-saliency detection has the potential to detect salient objects in various surveillance scenes collected by IoT devices. Figure 1 shows the schematic diagram of the co-saliency detection for cloud-based IoT applications. Various IoT devices in different locations collect the surveillance images. Then these images are transmitted to cloud servers for the subsequent co-saliency detection.

However, the existing co-saliency detection methods based on the deep artificial neural network (DNN) still suffer from two challenges, although they have been recognized to perform better than traditional hand-engineered image feature extractors [12], [15], [16], [17], [18]. The first challenge is to inefficiently represent the inter-image similarity and difference in the group images. It leads to the poor characterization of the co-saliency detection model to the common salient feature in the varied image backgrounds. The second challenge is to inefficiently extract the contextual information in the single image. It results in the difficulty of the co-saliency detection model to obtain the different-size receptive fields

This study was supported by the Project of Shenzhen International Cooperation Foundation (GJHZ20180926165402083), the Brazilian National Council for Research and Development (CNPq, Grant #304315/2017-6 and #430274/2018- 1), Science and Technology Planning Project of Guangdong Province, China (2018A050506031, 2019B010110001), Guangdong Natural Science Funds for Distinguished Young Scholar (2019B151502031), National Natural Science Foundation of China (61771464, U1801265), and the Fundamental Research Funds for the Central Universities. (*Corresponding authors: Chenchu Xu and Shuo Li*)

Z. Gao, C. Xu, and Shuo Li are with the Western University, Canada (e-mail: zgao246@uwo.ca; cxu332@uwo.ca; slishuo@gmail.com).

H. Zhang is with the School of Biomedical Engineering, Sun Yat-sen University, China (e-mail: zhangheye@mail.sysu.edu.cn).

V. H. C. de Albuquerque is with the Graduate Program in Applied Informatics, University of Fortaleza, Brazil (e-mail: victor.albuquerque@unifor.br).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2963701, IEEE Internet of Things Journal

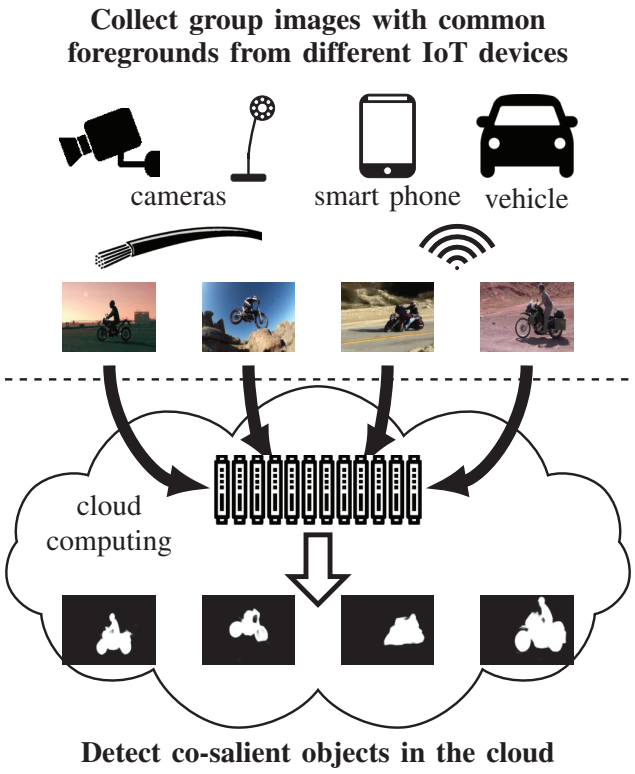

Fig. 1. The schematic diagram of the co-saliency detection for cloud-based IoT applications. The images are collected from varieties of IoT devices (e.g. surveillance camera, smart phone, and in-vehicle camera) in different locations. Then, they are transmitted to the cloud servers by wired or wireless communication, and gathered for co-saliency detection.

for perceiving varieties of objects.

In this paper, we propose a novel approach to detect co-salient objects in the trust IoT environment in order to address the inefficiency of inter-image information representation and single-image contextual information extraction. For improving the former efficiency, our approach builds a two-path information propagation scheme to simultaneously fuse the high-level information of common foreground in group images, and separate their difference information. For promoting the latter efficiency, our approach constructs a multi-stage context perception scheme to extract the semantic features of the different-size objects from the single image. Finally, our approach enables the stage-wise network refinement to help the network to learning from the enriched label knowledge.

Our contributions can be summarized as:

1. We enable the co-saliency detection in the surveillance scene analysis for trust IoT. It is better able to capture the common salient objects from surveillance data collected by IoT devices than the traditional saliency detection technologies.

2. Our multi-stage context perception strategy constructs an elaborately-designed neural network architecture to efficiently extract the semantic features with different receptive fields from the surveillance data of the single trustful IoT device.

3. Our two-path information propagation strategy provides a way to efficiently extract the similarity and difference information.between the surveillance data from diverse trustful IoT devices.

4. Our stage-wise network refinement enriches the knowledge for the neural network learning. It can help the neural network to learn the semantically common knowledge from the manually-labeled trustful IoT data.

This study has advantages to our preliminary conference paper [19]: (1) adding the stage-wise network refinement strategy to enrich the label knowledge for network learning; (2) constructing the two-path information propagation by refining the auto-encoder with the denoising scheme; (3) enriching the experimental analysis and discussion in order to better demonstrate the effectiveness and superiority of our approach.

II. Related Works

A. Trustful IoT in the smart city

Many researches have been proposed to safely merge the data from different IoT devices in the smart city [2], [5], [20], [21], [22], [23], [24], [25]. Li et al. [2] discuss the security issue in varieties of IoT applications and provide a policy-rule-based method to identify the untrustworthy and malicious nodes in IoT networks. Duan et al. [26] propose a game theoretic method for trust deviation. It can ensure the network security when minimizing the network latency and the energy consumption in order to decrease the overhead of networks during the trust derivation procedure. Adewuyi et al. [27] develop a dynamic model based on the application of collaborative download for trust management. It can address the challenges of trust parametrization, trust decay, and trust weighting. Wang et al. [28] introduce a cloud/edge architecture with adaptive service templates and trust evaluation. It can provide the balance dynamics between edge and cloud computing for solving the problems of increasing repeated service requirements and internal attacks. In the vehicular network, Rostamzadeh et al. [29] develop an information dissemination strategy with two modules: security check and message forward. It can move the trust concerns from cars to road segments for satisfying application-specific traffic requirements. Yang et al. [30] propose a block-chain-based strategy to construct a decentralized system for trust management. By the Bayesian inference method, this system can verify the information sent from nearby vehicles for addressing the challenges of message credibility. Jeong et al. [31] focus on the smart manufacturing system and provides a hierarchical model for trustful resource assignment.

B. Co-saliency detection in surveillance data analysis

Most co-saliency detection methods fall into three schemes based on bottom up, feature fusion and machine learning [14]. The bottom-up scheme aims to score the co-saliency degree of each pixel or region in the group images by hand-crafted co-saliency features, such as image contrast, pixel values and locations, gradient, etc. [15]. For instance, Li et al. [32] determine whether a pixel belongs to the co-saliency region by counting how many times it classified as a co-salient point among multiple predicted maps based on different image features. However, the bottom-up scheme highly rely on the user's experiences in the feature engineering, as well

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2963701, IEEE Internet of Things Journal

as the subjective tendency in the image observation. This characteristic leads that difficulty of the bottom-up methods to adapt to the varieties of practical image scenes. The feature-fusion-based scheme applies the existing co-saliency algorithms to compute the several predicted maps, and then fuse these maps to produce the final co-saliency map. In order to propose the appropriate map fusion method, the relationship among these predicted maps needs to be investigated. This relationship imposes the influence on the map fusion process as the prior constraint. For example, Cao et al. [33] consider the computation of the inter-image consistency information as a low-rank matrix recovery problem, i.e. minimize the norms of the low-rank matrix and the error matrix. After building the prior constraints of the matrix rank, the optimal error matrix are used to compute the fusion weights of the predicted maps. However, the performance of the fusion-based method is highly influenced by the co-saliency detection algorithms it uses. Its generalization may be weakened when the adopted co-saliency detection algorithms have the limited performance on the unseen group images. The learning-based scheme learns image feature representation of the co-saliency patterns in the group image. Early methods aim to construct the explicit learning model to approximate the map from group images to the co-saliency results [34]. For example, Wu et al. [35] cluster the multi-scale pre-segmented regions by feature bagging, and then apply the multiple kernel boosting to generate the co-saliency map. Similar to the bottom-up method, these methods however suffer from the insufficient feature representation ability of the hand-crafted image features. Benefit from the powerful feature representation of DNN [36], [37], [38], [39], some DNN-based methods have been developed to construct the implicit learning model for predicting the co-saliency map [40]. These methods focus on learning the feature representation from the auxiliary knowledge in another source domain, or the extract the common foreground feature in the group images [14].

III. Methodology

The surveillance scene analysis may facilitate the trust management of the IoT-based smart city. Our approach proposes an elaborately-designed architecture of the deep neural network to detect the co-salient objects from the images with varieties of natural scenes. The network architecture consists of three parts: multi-stage context perception, two-path information propagation, and stage-wise network refinement. Firstly, the multi-stage context perception proposes multiple inference block to produce different-size receptive field for capture the semantic information of objects with different sizes in varieties of scenes. It can efficiently extract the contextual information in the single scene. Then, the two-path information propagation has two information flows: one is to fuse the common semantic information in group images by the denoising convolutional auto-encoder (CAE), and the other one is to propagate the semantic information in every group image by single-channel propagation. It can efficiently represent the inter-image similarity and difference in group images. Finally, the stage-wise network refinement provides

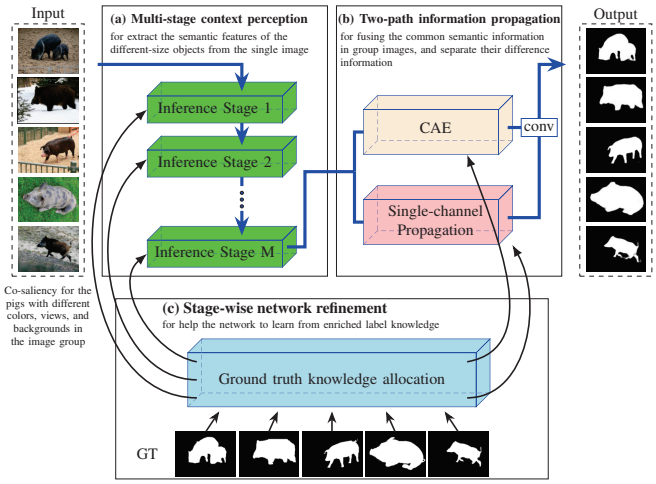

Fig. 2. The flowchart of our approach. It consists of three parts: multi-stage context perception, two-path information propagation and the stage-wise network refinement. (a) The multi-stage context perception extracts the semantic features of the different-size objects from the single image by the multi-stage inference block (the green cuboids). (b) The two-path information propagation contains CAE (the antique white cuboid) and single-channel propagation (the magenta cuboid), with the input as the output feature maps of the multi-stage context perception. The CAE fuses the common semantic information in group images. The single-channel propagation to separate the difference information in group images. (c) The stage-wise network refinement enriches the learning knowledge by adding multi-level representation of the ground truth information to the process of the multi-stage context perception and two-path information propagation. As the arrows shown, these information can facilitate the network learning as the label information of the intermediate network layers. "GT" is the ground truth.

the ground truth to supervise the multi-scale feature maps during the forward propagation of the other two parts of the network. It can enrich the learning knowledge to facilitate the semantic feature extraction. The schematic diagram of our approach is illustrated in Figure 2.

A. Multi-stage context perception for efficiently extracting the semantic feature in the single image

As shown in Figure 2, the raw group images are fed into the multiple inference stages in order to extract the semantic information. Denote the images in the group by $\{x_1, x_2, ..., x_N\}$. These images have similar semantic foregrounds but diverse backgrounds. The multi-stage context perception is defined as

$$z_j = f_j(x'_j; \theta_j), \quad j = 1, ..., M \tag{1}$$

where M is the number of the inference stages, f_j is the forward propagation of the jth inference stage, θ_j is the trainable parameters of f_j, x'_j and z_j are the input and output of f_j.

In Equation (1), for the ith raw image x_i in the group, x'_j is defined as

$$x'_j = \begin{cases} x_i & \text{if } j = 1 \\ z_{j-1} & \text{if } j > 1 \end{cases} \tag{2}$$

Every inference stage is constructed by the densely-connected network with dilated convolution. Denote the number of the convolution layers as K, the output feature map of the Kth layers is the output of the inference stage, i.e.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2963701, IEEE Internet of Things Journal

$z_j = q_K$. The dense connection of the different network layers in the inference stage can be formulated as [41]:

$$\begin{aligned} q_k &= C(p_k), \quad k = 1, 2, \cdots, K \\ p_k &= H(q_{k-1}, q_{k-2}, \ldots, q_1, x'_j) \end{aligned} \tag{3}$$

where p_k is the feature map as the input of the kth layer in the inference stage. $H(\cdot)$ is concatenation operation of the feature maps, i.e. combine multiple feature maps into a single feature map. $C(\cdot)$ is the dilated convolution [42]:

$$q_k = \sum_u \sum_v p_k(\beta \times u, \beta \times v) \cdot A(i - u, j - v) \tag{4}$$

where $p_k(i, j)$ is the pixel value of the feature map p_k at (i, j), A is the dilated convolution kernel with the dilated rate β, u and v are the coordinate offsets in A.

The above structure of the multi-stage context perception brings in two characteristics. First, the output of all convolution layers in an inference stage are directly connected with the output feature maps of this inference stage. This causes that the each inference stage can fuse the multiple feature representations which correspond to the diverse receptive field. Second, the output of the entire multi-stage context perception are directly connected with the output feature maps of all inference stages. This indicates that the multi-stage context perception can perceive different-size receptive fields, which have been obtained by intermediate convolution layers. In addition, the dilated convolution is capable to enlarge the receptive field because its convolution kernel is larger than the standard convolution according to Equation (4).

B. Two-path information propagation for efficiently extracting the inter-image similarity and difference in the group images

The two-path information propagation proposes two parallel feature representation paths. Its input is the concatenation of feature maps extracted from every single image by the multi-stage context perception. The first path aims to learn the inter-image similarity from the semantic feature maps of different images. It applies the denoising CAE to encode the concatenated feature maps with inter-image difference into a latent-space representation, and then reconstructs the output from this representation. For learning the inter-image similarity from this encode-decode process, this scheme detects the co-salient object in each of the group images as a single task, and consider this process as a multi-task learning. By feeding the label information of all group images as the feedback of this encode-decode process, This path can perceive the common co-salient foregrounds in the group images. The second path aims to extract the inter-image difference information belonging to each image in the group. To achieve this goal, it applies the channel splitting scheme to separate the input feature map along the channel direction. Then, it represents the separated features by the one-layer convolution. Finally, it fuses each channel with the output of CAE by the convolution layer. The fused map is the final output of the entire network.

Algorithm 1: Training process of the proposed approach

Input: group images x, label y, image number in the group N, number of inference stages M, epoch number K_1, group number K_2, learning rate r, momentum β, object function G

Output: Network parameter Θ after training

```
Initialize Θ by {θ_1^(0), θ_2^(0), ..., θ_M^(0), φ^(0), ψ^(0)};
for n_1 = 1 to K_1 do
    for n_1 = 1 to K_2 do
        for n_1 = 1 to K_2 do
            begin Forward Propagation
                1. Compute the output z of the
                   multi-stage context perception from x
                   according to Section III-A;
                2. Compute the output s_φ and s_ψ of the
                   two-path information propagation from
                   z according to Section III-B;
                3. Compute the objective function G in
                   Equation (5);
            end
            Backward Propagation;
        end
    end
end
```

C. Stage-wise network refinement for learning from the enriched label knowledge

The stage-wise network refinement enables the backpropagation from different stages of the entire network, inspired by the deep supervision [43], [44]. The backpropagation from the final output of the network is difficult to sufficiently train its aforementioned stages, because the these aforementioned stages are less sensitive to the variation of the error between the predicted co-saliency maps and the labels. In contrast, the proposed refinement enables the backpropagation from different stages of the network. It can enrich the feedback information which the network can learn from the label information. Specifically, it provide the feedback information to each inference stages in the multi-stage context perception, and the CAE and single-channel propagation in the two-path information propagation, shown as black arrows in Figure 2. This refinement process can be represented as the network training based on a specific-designed objective function G, i.e. the purpose of the network training is to minimize this objective function:

$$\Theta^* = \arg\min_{\Theta} G(y; \Theta) \tag{5}$$

where y is the label maps of the group images, which contains the real location of salient objects within the images. Θ is the set of the network parameters. G is the summation of the overall loss G_o of the network, and the summed loss G_c of all stages of the network:

$$\begin{aligned} G &= G_o(y; \Theta) + G_c(y; \Theta) \\ \Theta &= \{\theta_1, \theta_2, \ldots, \theta_M, \phi, \psi\} \end{aligned} \tag{6}$$

where $\theta_1, \theta_2, \ldots, \theta_M$ are the parameters of M inference stages shown in Equation (1) and Figure 2(a). ϕ is the parameters of CAE and ψ is the parameters of the single-channel propagation (i.e. the convolution kernel after the channel splitting) shown in Figure 2(b). G_c is defined by

$$G_c = L_1(s_\phi, y; \phi) + L_2(s_\psi, y; \psi) + \sum_{j=1}^{M} L_3(z_j, y; \theta_j) \quad (7)$$

where s_ϕ is the output of CAE and s_ψ is the output of the single-channel propagation. z_j is the output of the inference stage defined in Equation (1). L_1, L_2, L_3 are the loss functions to measure the error between the predicted results and the label, defined in Section III-D.

D. Implementation details

All input are resized to 256×256 pixels in both training and testing phases. In the multi-stage context perception, the number of the inference stage M is five. In every inference stage, the number of the convolution layers K are five. The dilated rates β of the five inference stages are 2, 2, 2, 4, 8 in order. The pooling operation is inserted between every two connected inference stages. In the two-path information propagation, the number of the convolution layers in the encoding and decoding process are both five. In the stage-wise network refinement, we propose a loss function $L = L_1 = L_2 = L_3$ to optimize the our co-saliency detection model by measuring the error between the predicted co-saliency region and the ground truth. L is defined as the summation of three losses: the mean absolute error, pixel-weighted cross-entropy [45], and the generalized Dice index [46]. In the training phase, the stochastic gradient descent is applied to optimize the entire neural network. The momentum is set 0.9. The weight decay is 1×10^{-4}. The learning rate is the 2×10^{-5}. Totally 80 epochs are used in the training phase (i.e. all samples in the training dataset are used for training by repeating 80 times). The training and testing are both implemented by TensorFlow (an open source library for machine learning). The pseudo code of the training phase in our approach has been shown in Algorithm 1, where the back-propagation indicates the update processing of the neural network.

IV. Experiments and Results

In this section, we have shown the effectiveness of our approach on three public datasets and its superiority to four state-of-the-art methods, as well as the implementation details and evaluations indices.

A. Experiment Setup

1) Datasets: Three public computer vision datasets have been applied to evaluate the performance of the co-saliency detection method.

Cosal [47]: This dataset has 2015 images divided into 50 image subsets, where the images are obtained from ILSVRC dataset and an online video set. Its characteristics include the dataset size is much larger than the other two co-saliency detection datasets, as well as the image backgrounds are highly diverse.

iCoseg [48]: This dataset has 643 images divided into 38 image groups. Its characteristics include the complex image backgrounds and multiple co-salient objects in the single image.

MSRC [14]: This dataset has 240 images divided into eight image groups (horse, tree, building, airplane, face, bicycle, car, grass). We do not use the group "grass" because no co-salient region appears in this group. Its characteristic is the co-salient objects have diverse shapes and color appearance.

2) Comparative methods: Our approach have been compared with four co-saliency detection studies at state of the art.

CBCS [15]: The CBCS develops a cluster-based algorithm to preserve the global relatedness in different images, and then applies three bottom-up cues to compute the final co-saliency maps.

CBCS_S [15]: The CBCS_S is an advanced version of CBCS by adding the group interactions proposed by [49].

CSHS [50]: The CSHS proposes the coarse saliency map based on the image border connectivity, and the fine saliency map based on the regional similarities of region pairs and the regional contrast of single region. Then, the co-saliency map is derived by computing the global similarity, as well as two maps with the inter-saliency and the object prior.

CSDW [47]: The CSDW builds two insights for co-saliency detection. (1) Transfer high-level feature representation by adding adaptive layers to convolutional network for improving the ability to extract the semantic information of co-salient objects. (2) Suppress the disturbance of the common background regions by using the neighborhood relatedness between image groups.

3) Evaluation indices: We evaluate the proposed approach and the four comparative methods by comparing the detected co-saliency map in every image and the ground truth. The ground truth of all images have been provided by the corresponding datasets. To quantify this comparison, we apply four evaluation indices: precision, recall, F-measure and average precision (AP) score.

The precision and recall are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

where TP is value of true positive which measures the number of pixels within in the co-saliency region correctly estimated. FP is the value of false positive which measures he number of pixels within in the co-saliency region incorrectly estimated. FN is the value of false negative which measures he number of pixels out of the co-saliency region correctly estimated.

Then the F-measure is the harmonic mean of the precision and recall to measure the accuracy. It is formulated as

$$\text{F-measure} = \frac{(\alpha^2 + 1) \times R \times P}{\text{R} + \alpha^2 \times \text{P}} \quad (9)$$

where R is recall, P is precision, and α is the weight parameter set by $\sqrt{0.3}$ in most state-of-the-art literature studies [14], [47], [51].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2963701, IEEE Internet of Things Journal

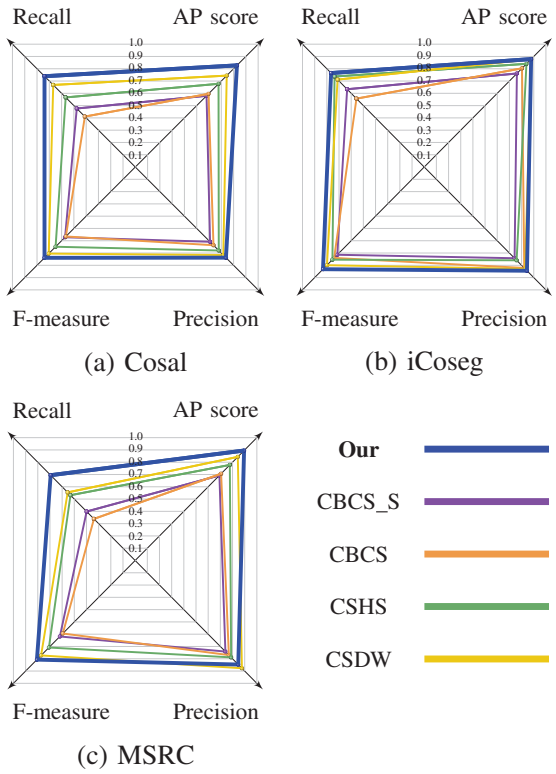

Fig. 3. Better performance of our approach than the comparative co-saliency detection methods evaluated by four evaluation indices: precision, recall, F-measure, and AP score. The three radar charts correspond to the performance on three datasets (Cosal, iCoseg and MSRC), respectively. The contour line displays the values ranged from 0 to 1. The closed curves with different colors represent different methods. The color legend is at the lower right corner of this figure. In these charts, the curves of our approach stay outermost, and thus signifies the highest values of the four indices. This indicates that our approach (blue) performs better than CBCS_S (purple), CBCS (orange), CSHS (green), and CSDW (gold).

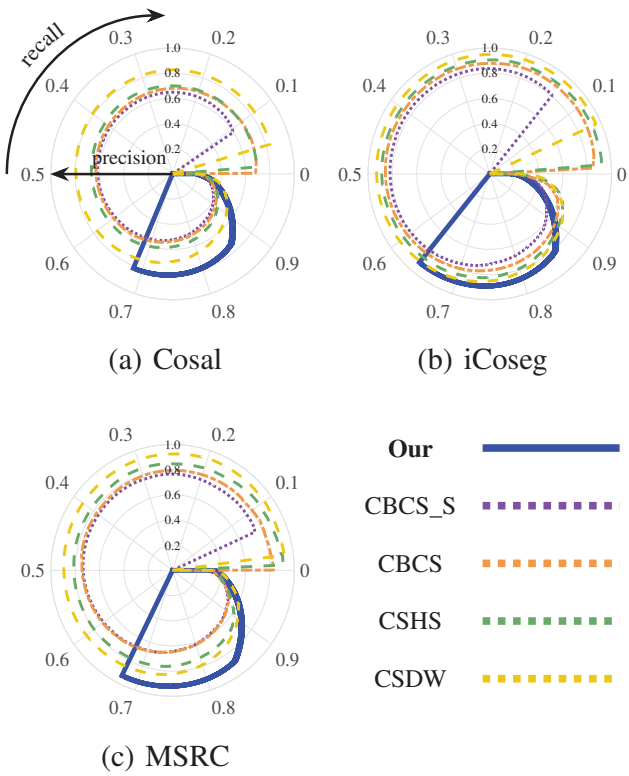

Fig. 4. Better performance of our approach than the comparative co-saliency detection methods evaluated by performance curve of the precision and recall. The three Nightingale rose charts correspond to the performance on three datasets (Cosal, iCoseg and MSRC), respectively. The radial direction represents the value of the precision ranged from 0 to 1, and the circumferential direction represents the value of the recall ranged from 0 to 1. The closed curves with different colors represent different methods. The color legend is at the lower right corner of this figure. In these charts, the curves of our approach fall in the bottom right part, while the curves of the comparative methods surround most of the circumference. Thus, the overall values of precision and recall in our approach is higher than the comparative methods. This indicates that our approach (blue) performs better than CBCS_S (purple), CBCS (orange), CSHS (green), and CSDW (gold).

Finally, the AP score is defined as the average values of precision for the recall value ranged between 0 to 1.

B. Outperformance over the state-of-the-art methods

Figure 3 displays the radar chart to show that our approach performs better in four evaluation indices (precision, recall, F-meausre and AP score) than the state-of-the-art methods. Each radar charts corresponds to the results computed in a dataset. In the radar chart, the vertex represents the evaluation index, the colored closed curves correspond to the performance of the comparative methods, the contour line displays the values ranged from 0 to 1. The results show that the performance curve of our approach (blue curve) stays outermost in all of the three radar charts. Thus, the performance curve of our approach has the larger area than those of CBCS (orange color), CBCS_S (purple color), CSHS (green color), and CSDW (gold color).

Figure 4 displays the Nightingale rose chart [52] to show that our approach gains better precision-recall performance curve. Each rose charts corresponds to the results computed in a dataset. In the rose chart, the radial direction represents the value of the precision ranged from 0 to 1, the circumferential direction represents the value of the recall ranged from 0 to 1, and the colored closed curves correspond to the performance of the comparative methods. The results show that the values of the precision and recall in our approach (blue color) fall in the bottom right part of the rose chart. In contrast, the curves of comparative methods surround most of the circumference of the rose charts. This indicates that the overall values of precision and recall in our apporach is higher than the comparative methods.

C. Better performance in the representative results

Figure 5 displays some representative results to show the outperformance of our approach in six aspects:

(1) Diverse structures of the same-class objects. Figure 5(a) shows the cargo ship and cruise ship as the example, which presents very different appearance and structures in the images. Our approach can obtain the much better co-saliency maps of the two ships than the comparative methods with respect to the ground truth (GT).

(2) Fine structures. Figure 5(b) shows two deer as the example. They have significant fine structure with respect to their main bodies, i.e. the antlers and tails. Our approach is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2019.2963701, IEEE Internet of Things Journal

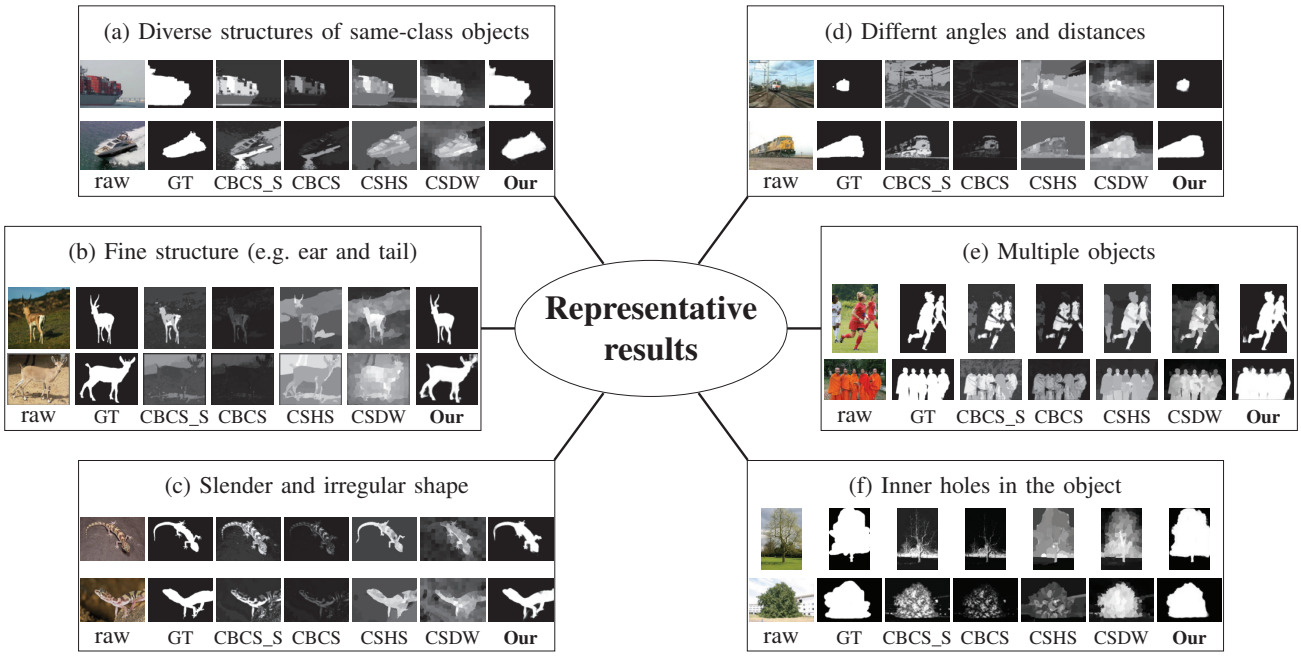

Fig. 5. Representative results to show the better performance of our framework than the state-of-the-art approaches in six aspects: (a) Able to capture the diverse structures of the same-class objects (e.g. cargo ship and cruise ship). (b) Capable to perceive fine structure of the objects (e.g. antlers and tails of the deer). (c) Able to obtain more complete structure when the object has slender and irregular shape (e.g. lizard). (d) Capable to detect the object in different shooting angles and distances under diverse backgrounds (e.g. train). (e) Able to extract the co-saliency map with multiple objects in a scene (e.g. players in the soccer game and the gathering monks). (f) Robust to the disturbance of the inner holes of the co-salient objects (e.g. tree).

able to perceive these detailed parts of the deer while the comparative methods cannot.

(3) Slender and irregular shape. Figure 5(c) displays the images of two lizards. The lizard has the slender body with the irregular shape depending on their postures. Comparing with the comparative methods, our approach is capable to the capture the more complete body of the lizard.

(4) Different angles and distances. Figure 5(d) gives the example of the train images with two shooting angles and distances. This causes that the trains present diverse appearance and shapes in the images. Our approach has the ability to accurately detect the trains as the foreground under the disturbance of diverse background, e.g. utility poles and rails.

(5) Multiple objects. Figure 5(e) shows the two scenes with multiple persons: soccer game and monk gathering. The comparative methods cannot well perceive all co-salient persons, as well as distinguish between them and the background. However, our approach performs well.

(6) Inner holes in the object. Figure 5(f) displays the case that the co-saliency object whose morphology has many inner holes. The tree is an example. The results show that our approach has much better robustness to these inner holes than the comparative methods.

V. Conclusion

Trustful IoT network is crucial in the smart city. The numerous information from surveillance devices may provide key information to help the trust management of IoT. In this paper, we have proposed a novel co-saliency detection approach for surveillance scene analysis based on an elaborately-designed architecture of the deep neural network, including the multi-stage context perception network, two-path information propagation network, as well as stage-wise network refinement strategy. The extensive experiments performed on three public datasets demonstrate that effectiveness of our approach, as well as its superiority to the four state-of-the-art approaches. In addition, our approach needs to determine the size of the image group in advance, as well as invalid in video data. To address the two limitations, our future study will consider the size-varied group as the network input, as well as be improved based video data to investigate the long-time co-saliency relationship in the diverse surveillance IoT devices.

References

[1] Ş. Kolozali, M. Bermudez-Edo, N. Farajidavar, P. Barnaghi, F. Gao, M. I. Ali, A. Mileo, M. Fischer, T. Iggena, D. Kuemper, and R. Tonjes, "Observing the pulse of a city: A smart city framework for real-time discovery, federation, and aggregation of data streams," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2651–2668, 2019.

[2] W. Li, H. Song, and F. Zeng, "Policy-based secure and trustworthy sensing for internet of things in smart cities," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 716–723, 2018.

[3] M. Frustaci, P. Pace, G. Aloi, and G. Fortino, "Evaluating critical security issues of the IoT world: Present and future challenges," *IEEE Internet of Things Journal*, vol. 5, no. 4, pp. 2483–2495, 2018.

[4] J. Guo, I. R. Chen, and J. J. P. Tsai, "A survey of trust computation models for service management in internet of things systems," *Computer Communications*, vol. 97, pp. 1–14, 2017.

[5] J. Hou, L. Qu, and W. Shi, "A survey on internet of things security from data perspectives," *Computer Networks*, vol. 148, pp. 295–306, 2019.

[6] R. Leyva, V. Sanchez, and C. T. Li, "Video anomaly detection with compact feature sets for online performance," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3463–3478, 2017.

[7] M. U. K. Khan, H. S. Park, and C. M. Kyung, "Rejecting motion outliers for efficient crowd anomaly detection," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 2, pp. 541–556, 2019.

[8] H. Kim, L. Mokdad, and J. Ben-Othman, "Designing uav surveillance frameworks for smart city and extensive ocean with differential perspectives," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 98–104, 2018.

[9] J. Wang, Y. Wang, D. Zhang, Q. Lv, and C. Chen, "Crowd-powered sensing and actuation in smart cities: Current issues and future directions," *IEEE Wireless Communications*, vol. 26, no. 2, pp. 86–92, 2019.

[10] I. Ahmed, A. Ahmad, F. Piccialli, A. K. Sangaiah, and G. Jeon, "A robust features-based person tracker for overhead views in industrial environment," *IEEE Internet of Things Journal*, vol. 5, no. 3, pp. 1598–1605, 2018.

[11] J. Huang, N. Duan, P. Ji, C. Ma, feng hu, Y. Ding, Y. Yu, Q. Zhou, and W. Sun, "A crowdsource-based sensing system for monitoring fine-grained air quality in urban environments," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3240–3247, 2019.

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

[13] H. Li, F. Meng, and K. N. Ngan, "Co-salient object detection from multiple images," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1896–1909, 2013.

[14] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms: Fundamentals, applications, and challenges," *ACM Transactions on Intelligent Systems and Technology*, vol. 9, no. 4, pp. 38:1–31, 2018.

[15] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Transactions on Image Processing*, vol. 22, no. 10, pp. 3766–3778, 2013.

[16] Z. Gao, J. Chung, M. Abdelrazek, S. Leung, W. K. Hau, Z. Xian, H. Zhang, and S. Li, "Privileged modality distillation for vessel border detection in intracoronary imaging," *IEEE Transcations on Medical Imaging*, 2019.

[17] Z. Gao, X. Wang, S. Sun, D. Wu, J. Bai, Y. Yin, X. Liu, H. Zhang, and V. H. C. de Albuquerquee, "Learning physical properties in complex visual scenes: An intelligent machine for perceiving blood flow dynamics from static ct angiography imaging," *Neural Networks*, 2019.

[18] Z. Gao, H. Xiong, X. Liu, H. Zhang, D. Ghista, W. Wu, and S. Li, "Robust estimation of carotid artery wall motion using the elasticity-based state-space approach," *Medical Image Analysis*, vol. 37, pp. 1–21, 2017.

[19] M. Li, S. Dong, K. Zhang, Z. Gao, X. Wu, H. Zhang, G. Yang, and S. Li, "Deep learning intra-image and inter-images features for co-saliency detection," in *British Machine Vision Conference (BMVC)*, 2018.

[20] L. Wu, X. Du, M. Guizani, and A. Mohamed, "Access control schemes for implantable medical devices: A survey," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1272–1283, 2017.

[21] Z. Guan, J. Li, L. Wu, Y. Zhang, J. Wu, and X. Du, "Achieving efficient and secure data acquisition for cloud-supported internet of things in smart grid," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1934–1944, 2017.

[22] M. Shen, Y. Deng, L. Zhu, X. Du, and N. Guizani, "Privacy-preserving image retrieval for medical IoT systems: A blockchain-based approach," *IEEE Network*, vol. 33, no. 5, pp. 27–33, 2019.

[23] M. Shen, X. Tang, L. Zhu, X. Du, and M. Guizani, "Privacy-preserving support vector machine training over blockchain-based encrypted IoT data in smart cities," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 7702–7712, 2019.

[24] C. Wang, S. Dong, X. Zhao, G. Papanastasiou, H. Zhang, and G. Yang, "Saliencygan: Deep learning semi-supervised salient object detection in the fog of IoT," *IEEE Transactions on Industrial Informatics*, 2019.

[25] Z. Gao, H. Zhang, S. Dong, S. Sun, X. Wang, G. Yang, W. Wu, S. Li, and V. H. C. de Albuquerque, "Salient object detection in the distributed cloud-edge intelligent network," *IEEE Network*, 2019.

[26] J. Duan, D. Gao, D. Yang, C. H. Foh, and H. H. Chen, "An energy-aware trust derivation scheme with game theoretic approach in wireless sensor networks for IoT applications," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 58–69, 2014.

[27] A. A. Adewuyi, H. Cheng, Q. Shi, J. Cao, A. MacDermott, and X. Wang, "CTRUST: A dynamic trust model for collaborative applications in the internet of things," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5432–5445, 2019.

[28] T. Wang, G. Zhang, A. Liu, M. Z. A. Bhuiyan, and Q. Jin, "A secure IoT service architecture with an efficient balance dynamics based on cloud and edge computing," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4831–4843, 2019.

[29] K. Rostamzadeh, H. Nicanfar, N. Torabi, S. Gopalakrishnan, and V. C. M. Leung, "A context-aware trust-based information dissemination framework for vehicular networks," *IEEE Internet of Things Journal*, vol. 2, no. 2, pp. 121–132, 2015.

[30] Z. Yang, K. Yang, L. Lei, K. Zheng, and V. C. M. Leung, "Blockchain-based decentralized trust management in vehicular networks," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1495–1505, 2019.

[31] S. Jeong, W. Na, J. Kim, and S. Cho, "Internet of things for smart manufacturing system: Trust issues in resource allocation," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4418–4427, 2018.

[32] Y. Li, K. Fu, Z. Liu, and J. Yang, "Efficient saliency-model-guided visual co-saliency detection," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 588–592, 2015.

[33] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 4175–4186, 2014.

[34] H. Yu, K. Zheng, J. Fang, H. Guo, W. Feng, and SongWang, "Co-saliency detection within a single image," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018, pp. 7509–7516.

[35] L. Wu, Z. Liu, H. Song, and O. L. Meur, "RGBD co-saliency detection via multiple kernel boosting and fusion," *Multimedia Tools and Applications*, vol. 77, no. 16, pp. 21 185–21 199, 2018.

[36] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[37] Z. Gao, S. Wu, Z. Liu, J. Luo, H. Zhang, M. Gong, and S. Li, "Learning the implicit strain reconstruction in ultrasound elastography using privileged information," *Medical Image Analysis*, 2019.

[38] Z. Gao, Y. Li, Y. Sun, J. Yang, H. Xiong, H. Zhang, X. Liu, W. Wu, D. Liang, and S. Li, "Motion tracking of the carotid artery wall from ultrasound image sequences: A nonlinear state-space approach," *IEEE Transcations on Medical Imaging*, vol. 37, no. 1, pp. 273–283, 2018.

[39] H. Zhang, Z. Gao, L. Xu, X. Yu, K. C. L. Wong, H. Liu, L. Zhuang, and P. Shi, "A meshfree representation for cardiac medical image computing," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 6, no. 1800212, 2018.

[40] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, and J. Han, "A self-paced multiple-instance learning framework for co-saliency detection," in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 594–602.

[41] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[42] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representation (ICLR)*, 2016.

[43] C. Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015, pp. 562–570.

[44] L. Wang, C. Y. Lee, Z. Tu, and S. Lazebnik, "Training deeper convolutional networks with deep supervision," *Computing Research Repository (CoRR)*, vol. abs/1505.02496, 2015.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[46] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, 2011.

[47] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *International Journal of Computer Vision*, vol. 120, no. 2, pp. 215–232, 2016.

[48] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3169–3176.

[49] L. Wei, S. Zhao, O. E. F. Bourahla, X. Li, and F. Wu, "Group-wise deep co-saliency detection," in *International Joint Conferences on Artifical Intelligence (IJCAI)*, 2017, pp. 3041–3047.

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

[51] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 4039–4048.

[52] N. I. Fisher, *Statistical Analysis of Circular Data*. Cambridge University Press, 1995.