Generalised Co-Salient Object Detection

Jiawei Liu, Jing Zhang, Ruikai Cui, Kaihao Zhang, Weihao Li, Nick Barnes

Abstract—We propose a new setting that relaxes an assumption in the conventional Co-Salient Object Detection (CoSOD) setting by allowing the presence of “noisy images” which do not show the shared co-salient object. We call this new setting Generalised Co-Salient Object Detection (GCoSOD). We propose a novel random sampling based Generalised CoSOD Training (GCT) strategy to distill the awareness of inter-image absence of co-salient objects into CoSOD models. It employs a Diverse Sampling Self-Supervised Learning (DSSL) that, in addition to the provided supervised co-salient label, introduces additional self-supervised labels for noisy images (being null, that no co-salient object is present). Further, the random sampling process inherent in GCT enables the generation of a high-quality uncertainty map highlighting potential false-positive predictions at instance level. To evaluate the performance of CoSOD models under the GCoSOD setting, we propose two new testing datasets, namely CoCA-Common and CoCA-Zero, where a common salient object is partially present in the former and completely absent in the latter. Extensive experiments demonstrate that our proposed method significantly improves the performance of CoSOD models in terms of the performance under the GCoSOD setting as well as the model calibration degrees.

Index Terms—Generalised Co-Salient Object Detection, Co-Salient Object Detection Testing Dataset, Uncertainty Estimation, Model Calibration.

I. INTRODUCTION

Co-salient object detection (CoSOD) [1]–[4] aims to localise a common salient object in a group of images. Different from single image based salient object detection (SOD) [5]–[7], CoSOD performs group-wise salient object detection, and the resulting salient map of each image within the group should reflect the common salient attribute across the group. To this end, most of the existing CoSOD models focus on group-wise attention/correlation modeling, aiming to discover the common salient regions.

We observe that the conventional CoSOD setting assumes the presence of co-salient object in each individual image belonging to the same pre-sorted group. The existing CoSOD models, despite performing well in localising the co-salient object across a group of pre-sorted images, find it challenging to deal with the unsorted image group, where co-salient object is only partially present. As illustrated in Fig. 1, CADC [3], one of the SOTA models in conventional CoSOD benchmarks, has difficulty in acknowledging the absence of the co-salient object in the introduced “noisy image” due to the knowledge distilled from the conventional CoSOD training.

We propose a new setting that relaxes the conventional CoSOD setting by allowing the presence of varying numbers of “noisy images” that do not share the co-salient object of the pre-sorted group. We call this new setting as Generalised Co-Salient Object Detection (GCoSOD). It is motivated by a Photo Sorting problem that requires finding subgroups in photo collections that are each associated with a common topic/object without tags as prior knowledge. It is natural to assume that the entire photo collection does not have any common salient object that is universally present in each individual image and the presence ratio of common salient object in individual cases may vary. GCoSOD can be iteratively applied to the photo collections: (1) find the co-salient object of current iteration that has the most appearances in the collection; (2) relocate the retrieved images with the co-salient object to the corresponding subgroup.

To address the false-positive predictions of co-salient objects in the noisy images, we propose a generalised CoSOD training (GCT) strategy to distill the awareness of the potential absence of co-salient object in the inter-image level into CoSOD models. Instead of feeding the models with training images from the pre-sorted groups, GCT partially/wholly replaces them with images containing salient objects of distinct categories, regarded as “noisy images”. The “noisy images”, associated with complete-negative ground truth, can present at different levels ranging from 0 to 100%, leading to two extreme cases: 1) 0% of the images are noisy, which is the same as the conventional CoSOD setting, where co-salient object exists in each image of the group; and, 2) 100% of the images are noisy, indicating that there is no overlap in terms of the salient object categories, or there is no common salient object. We find that GCT not only improves the handling of “noisy images” by existing CoSOD models, but also significantly improves their model calibration degrees [8] (see Fig. 5 and Tab. II).

The core of GCT lies in the Diverse Sampling Self-

J. Liu, J. Zhang, R. Cui, K. Zhang, W. Li and N. Barnes are with the College of Engineering and Computer Science, The Australian National University, Canberra, ACT, Australia. E-mail: {Jiawei.Liu3, Jing.Zhang, Ruikai.Cui, Kaihao.Zhang, Weihao.Li, Nick.Barnes}@anu.edu.au

Fig. 1. The group has “Watermelon” as the common salient object shared across its images except for an introduced “noisy image” which contains a “cat” and a “butterfly”. Existing models, such as CADC [3], find it challenging to acknowledge the absence of a common salient object in the “noisy image”. On the contrary, our proposed GCT, implemented on top of CADC, successfully avoids false-positive predictions in the “noisy image”.
Supervised Learning (DS$^3$L) that randomly determines the noise level (the percentage of “noisy images”) and the categories of the “noisy images”. It associates each training image with diverse annotations: 1) with foreground areas highlighting the co-salient object when it exists; and, 2) with complete-negative ground truth when it is sampled as a “noisy image”. Further, the diverse ground truth co-saliency maps enabled by DS$^3$L make it possible to generate high-quality uncertainty maps highlighting potential false-positive predictions at an instance level (see Fig. 6 and Fig. 7).

In addition to the training strategy, we also observe that existing CoSOD testing datasets pre-sort images into groups, where a unique co-salient object is present universally in each individual image within each pre-sorted group. Such groups are unsuitable to evaluate the ability of CoSOD models to handle “noisy images” in the GCoSOD setting. We re-arrange the CoCA dataset [9] and propose two new testing datasets, namely CoCA-Common and CoCA-Zero, where a common salient object is partially present in the former and completely absent in the latter. Extensive experiments demonstrate the superiority of our method in the GCoSOD setting where the “noisy images” are present to various extents.

We summarise our main contributions as: 1) we propose a GCT strategy to enable the existing CoSOD models to handle “noisy images” in our newly proposed GCoSOD setting; 2) we present a new uncertainty estimation technique via diverse ground truth sampling, enabled by DS$^3$L; 3) we re-arrange the CoCA dataset [9] to produce two new CoSOD testing datasets, namely CoCA-Common and CoCA-Zero, to evaluate the performance of CoSOD models under the proposed GCoSOD setting where “noisy images” can present to various extents; 4) we perform extensive experiments to verify the effectiveness of the proposed GCT in improving existing CoSOD models in terms of “noisy image” handling and model calibration degree.

II. RELATED WORKS

Co-Salient Object Detection: Traditional CoSOD methods are based on handcrafted features [10]–[13]. These were succeeded by deep neural network (DNN) based approaches with more representative deep features. Early DNN based CoSOD models combine the representation power of convolutional neural networks (CNNs) with traditional methods [14]. Recent works explore the correlation of co-salient objects across the group of images with solely CNNs. They can be categorised into (1) affinity based methods [1], [3], [4], [15]–[20], and (2) graph neural network (GNN) based methods [2], [21]–[23].

Affinity based methods search for common features across a group of images based on their feature affinity to localise the co-salient objects. This has been achieved through pixel-level affinity [15], a recurrent co-attention unit [16], self-attention mechanism [3], Principal Component Analysis [1], [17], etc. CoADNet [18] develops a two-stage aggregation model, utilising a combination of a spatial attention mechanism and channel attention mechanism, together with a distribution mechanism to find the co-salient object in different image groups. [19] builds 4D correlation volumes on the Gromov-Wasserstein distance to discover co-salient object through pairwise pixel similarities. In addition to common features, [20] further employs saliency priors to eliminate distractions from similar background where common objects can potentially appear. Graph Neural Networks (GNN) based methods construct graph models on deep features and find common features through graph clustering or graph propagation. [21] designs a graph manifold ranking strategy to propagate the group common features to localise co-salient object in individual images. [2], [22] propose an adaptive graph to learn inter- and intra-image consistency. This idea is extended in [23] by leveraging multiple graph models in a unified framework.

In this paper, we propose a GCT to distill the awareness of potential inter-image absence of the common salient object into the existing CoSOD models, improving their robustness to the presence of “noisy images” that do not share the group-wise co-salient object.

CoSOD Datasets: Existing CoSOD testing datasets include MSRC [24], iCoSeg [25], CoSal2015 [26], CoSOD3k [1] and CoCA [9]. MSRC and iCoSeg contains 233 images divided into 7 groups and 643 images divided into 38 groups respectively. High appearance consistencies are observed in both foreground salient objects and background scenes of those images, making them relatively easy for CoSOD. CoSal2015 [26] and CoSOD3k [1] are two large-scale CoSOD datasets, comprising of 2,015 and 3,316 images in 160 and 80 groups respectively. Despite the increase in size and variance in both foreground objects and background scenes, most of the images constitute single foreground objects, which make them easily solvable with single RGB image based SOD models. This is addressed to some extent in the newly proposed CoCA [9] dataset where salient objects not belonging to the co-salient categories are introduced in each image as distractions.

Model Calibration: Guo et al. [8] discover that modern deep neural networks are prone to making over-confident predictions. To remedy this issue, [27], [28] generate adversarial samples [29] to increase the diversity of training dataset. Kuamr et al. [30] introduce a trainable model calibration error as a regularization term into their objective function. [31] investigates the effect of dataset shift on prediction accuracy and model calibration degree. Two main strategies have been widely studied to achieve model calibration, namely label relaxation [32] and temperature scaling [8]. Label relaxation aims to relax the supervision signals, thus generating smoothed labels [32] or disturbed labels [33]. ASLP [34] proposes an adaptive stochastic label perturbation following the Maximum Entropy Inference [35] of statistical mechanics. We propose a novel approach that augments training data with diversified self-supervised labels to reduce the overconfidence in falsely predicted co-salient objects.

Uncertainty Estimation: Uncertainty estimation aims to evaluate model prediction in terms of its reliability [36]. Approaches have been explored in several dense segmentation tasks [6], [37]–[39]. Existing techniques for uncertainty estimation usually adopt the Bayesian Neural Network [40], [41] or stochastic prediction network [42], [43]. The former focus on model uncertainty, indicating inadequate unawareness of the true model that generates the observed data, and latter is usually designed to model data uncertainty, representing the inherent labeling noise. In this paper, we introduce a
Fig. 2. Samples from the proposed “CoCA-Common” dataset (top) and “CoCA-Zero” dataset (bottom). The “Candle” group of CoCA-Common has candle as the co-salient object shared across the primary images. It also includes noisy images whose salient objects have categories different from candle. On the other hand, the group of CoCA-Zero consists of images with distinguished salient object categories from each other.

III. OUR METHOD

A. Overview

For co-salient object detection, we have a dataset \( D = \{G_k\}_{k=1}^K \), consisting of \( K \) groups of images with the corresponding ground truth saliency maps, which is \( G_k = \{x_{kj}, y_{kj}\}_{j=1}^{N_k} \), where \( x_{kj} \) and \( y_{kj} \) indicate an image-groundtruth pair of index \( j \) from the \( k^{th} \) group, and \( N_k \) is the size of group \( G_k \). Given the training dataset \( D \), the typical training procedure of existing CoSOD models performs co-salient object detection within each pre-sorted group, where a common salient object is present in each individual image of the same group. Despite achieving strong performance under the conventional CoSOD setting, existing CoSOD models are not robust to the introduction of “noisy images” which do not share the common salient object.

To avoid false-positive predictions for the “noisy images”, we propose a Generalised CoSOD Training strategy that allows the model to detect the absence of common salient objects in “noisy images”. In addition, to evaluate the performance of CoSOD models under the proposed GCoSOD setting, we introduce two new testing datasets, CoCA-Common and CoCA-Zero, which are generated by re-arranging the CoCA [9] testing dataset, with samples presented in Fig. 2. The former contains groups where only a fraction of the images contain the common salient object, and the latter contains groups of images with no co-salient object at all.

B. Generalised CoSOD Training Strategy

We propose a random sampling based Generalised CoSOD Training strategy, that enables CoSOD models to deal with scenarios where there is complete or partial absence of the common salient object across the images of a group. Firstly, instead of requiring a common salient object to be present in each individual image of a group, we relax this requirement by defining the co-salient object as a salient object with the most diverse sampling based uncertainty estimation method that is capable of highlighting potential false-positive predictions at an instance level.
appearances in the images of the same group. The proposed
GCT randomly introduces “noisy images” at various ratios. It
further employs a Diverse Sampling Self-Supervised Learning
(DS^{3}L) to introduce additional self-supervised labels, with
complete 0 indicating full background, for images drawn as
“noisy images” with no common salient object. With this, for
each pre-sorted co-saliency group, we define it as a “primary
group”, associated with a corresponding “secondary group”
containing images that do not share the co-saliency object.

**Diverse Sampling Self-Supervised Learning (DS^{3}L):** We
randomly draw images from the secondary groups as “noisy
images” to partially replace images in the primary group. The
noisy-image-replacement process: (1) randomly determines the
replacement ratio by sampling from a uniform distribution
as: $\hat{r} \sim U(0, 1)$, and computes the number of replacement
images as $r = \lfloor N_i \times \hat{r} \rfloor$, where $\lfloor \cdot \rfloor$ is a floor function; (2) We
randomly select $r$ groups from the $K - 1$ secondary groups and
draw a single noisy image from each of them to replace the
images in the primary group. The stochastic attribute imbued
in the random selection of noisy image distills the awareness
of potential absence of co-saliency object in some noisy inputs,
which could contain non-co-saliency object of various classes,
textures and shapes.

**Extreme Cases:** Two extreme cases exist in our random
sampling based training dataset preparation process: 1) the
primary group is unchanged with proportion of the secondary
groups equal to 0; and 2) the primary group is totally replaced
with secondary groups, where no common salient objects
exists across the new group. With the first case, our training
dataset is the same as the conventional CoSOD setting, where
common salient objects exist across the group of images. The
second case is designed to expose the model to scenarios
where the proportion of images containing the common salient
object is below certain threshold. For most of the cases, our
random sampling strategy generates new groups with only
a subset of images containing no common salient objects,
which is consistent with real-life scenarios (i.e., an imperfectly
curated set).

**C. Uncertainty Estimation via Random Sampling**

The “random sampling” process inherent to the Generalised
CoSOD Training creates multiple annotations for each training
image: 1) when the image belongs to the primary group, its
ground truth includes foreground areas covering the co-saliency
object; 2) when the image belongs to the secondary group, its
ground truth is completely negative, indicating the absence of
the co-saliency object. The divergence modelling, conditioned
on the group context, improves the model awareness of the
conditional occurrence of a co-saliency object. It enables the
model to produce less confident predictions in difficult cases
where the noisy salient object demonstrates certain appear-
ance similarity to the common salient object. We define the
uncertainty as the entropy of the prediction:

$$u_{ki} = \mathbb{H}(p_{\theta}(\hat{y}_{ki} | \{x_{kj}\}_{j=1}^{N_k}))$$
$$= -p_{\theta}(\hat{y}_{ki} | \{x_{kj}\}_{j=1}^{N_k}) \log(p_{\theta}(\hat{y}_{ki} | \{x_{kj}\}_{j=1}^{N_k}) + \epsilon)$$

(1)

where $p_{\theta}(\hat{y}_{ki} | \{x_{kj}\}_{j=1}^{N_k}) \in \mathbb{R}^{H \times W}$ is a normalised conditional
co-saliency prediction with $H$ and $W$ representing the original
image sizes, $i$ and $j$ indexes the samples in group $k$, and $k$
indexes the co-saliency groups, $\epsilon = 10^{-6}$ is introduced to
avoid the $\log(0)$ error.

**D. New Testing Datasets for GCosOD Setting**

We propose two new CoSOD testing datasets, namely
“CoSOD-Common” and “CoSOD-Zero”, to measure the abil-
ity of CoSOD models to handle “noisy images” that do not
share the common salient object. These two datasets are built
by re-arranging the CoCA [9] dataset.
1) CoCA-Common: CoCA-Common contains 177 groups of images where each group consist of images only some of which show a co-salient object. There are 59 distinct groups, each having three different co-salient object presence ratios falling in the ranges of \([0.2, 0.4), [0.4, 0.6)\) and \([0.6, 0.8]\) respectively. The overall presence ratio of co-salient objects is between 0.2 and 0.8. Fig. 3 presents the co-salient object presence ratios of CoCA-Common. The CoCA-Common dataset is designed to measure the robustness of CoSOD models in scenarios where “noisy images” are present in different scales.

2) CoCA-Zero: CoCA-Zero contains 55 groups of images, where each group may have a different number of images, and there is no common salient object across the images of each group. These groups are formed by selecting at most one image per group from the CoCA [9] dataset. Visual examination is applied as post-processing, ensuring no accidental occurrence of co-salient object in each group. For example, the “Person” category has a 78% presence in the “Basketball” group of CoCA [9], making it an unwanted, yet commonly presented secondary salient object. CoCA-Zero is designed to measure the model behaviour in the scenarios where no common salient object exists.

E. Evaluation

With the two new testing datasets, we aim to produce \(0\) (full background) for images containing no co-salient object. We find that widely used CoSOD evaluation metrics (F-measure \(F_\beta\), S-measure [44] \(S\) and E-measure [45] \(\xi\)) are unable to evaluate CoSOD models under the GCoSOD setting. For an image with \(0\) as ground truth, as the true positive is 0, we then have precision \(= 0\), and the resulting F-measure \(F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}\) is also 0, which is independent of the quality of the predictions. Structure measure (S-measure) [44] evaluates the structure similarity between the prediction and ground truth in both region and object levels. Similar to the F-measure, based on the region and object levels mean foreground predictions \(\hat{s}\), the two corresponding structure similarities \((S_r, S_o)\) for region and object level similarity respectively) are \(0\) for all predictions, making S-measure ineffective in measuring model prediction in our scenario. The Enhanced-alignment measure (E-measure) [45] computes a bias matrix for the prediction \((\varphi_s)\) and the ground truth map \((\varphi_y)\) respectively, which is the distance between each pixel within the prediction or the ground truth and its global mean. The final alignment matrix \(\xi\) is then defined as 
\[
\xi = \frac{\varphi_s \odot \varphi_y}{\sum \varphi_s \odot \varphi_y},
\]
where \(\odot\) is the Hadamard product. Given accurate predictions with \(0\) as ground truth, we have \(\varphi_s = 0\), leading to \(\xi = 0\) in this case.

Considering that the conventional metrics are less effective for evaluating the model performance under the proposed GCoSOD setting, we adopt the widely used mean Intersection-over-Union (mIoU) metrics. For our case with \(0\) as ground truth, the resulting mIoU = 1 for accurate predictions is consistent with the quality of the prediction.

IV. EXPERIMENTAL RESULTS

A. Settings

1) Implementation Details: We adopt CADC [3] as our baseline to implement the proposed GCT (GCT-CADC). The model is initialised with ImageNet pretrained weights for the VGG-16 backbone. We keep other settings of the original work unchanged by setting the training size to 256 × 256 and batch size to 14. For data augmentation, we employ random flipping and random cropping. The model is trained with the Stochastic Gradient Descent (SGD) optimiser for 40,000 iterations. The initial learning rate is set to 0.01 and then halved at the \(20,000^{th}\) and the \(40,000^{th}\) iteration. We perform all ablation study experiments on GCT-CADC and interchangeably term it as “Ours”. We also generalise our proposed GCT onto ICNet [20] and DCFM [4] to produce GCT-ICNet and GCT-DCFM and show the results in Tab. I.

2) Datasets: Following the baseline model [3], we use a subset of COCO dataset [46], containing 9,213 images in 65 groups, and a synthetic dataset [3] for training. We use our proposed datasets, CoSOD-Common and CoSOD-Zero, to test the generalisation ability of existing CoSOD models. Further, we use our proposed datasets, CoCA-Common and CoCA-Zero, to test the performance of the existing CoSOD models in the GCoSOD setting, and three widely used CoSOD datasets, including CoSal2015 [26], CoSOD3k [1] and CoCA [9], to evaluate our method under the conventional CoSOD setting.

3) Co-saliency Metrics: We adopt the mean Intersection-over-Union (mIoU) and Mean Absolute Error (MAE) to evaluate the performance in GCoSOD setting, and the maximum F-measure \((F_\beta^\text{max})\), Structure-measure \((S_\alpha)\) [44] and maximum E-measure \((E_\xi^\text{max})\) [45] and MAE for evaluation in the conventional CoSOD setting.

4) Calibration Metrics: We adopt two calibration measures to measure the model calibration degree including Expected Calibration Error (ECE) [47] and reliability diagrams [48], [49]. We adopt the equal-width ECE [50] defined as:
\[
\text{ECE} = \frac{K}{N} \sum_{i=1}^{K} \frac{|B_i|}{N} \left(\text{acc}(B_i) - \text{conf}(B_i)\right)
\]
where predictions are grouped into \(K\) bins and \(\text{acc}(B_i)\) and \(\text{conf}(B_i)\) represent mean accuracy and mean confidence of predictions belonging to the \(i^{th}\) bin, \(|B_i|\) is the size of the \(i^{th}\) bin and \(N\) is the total number of predictions. Here we set the number of bins to \(K = 10\). Reliability diagram plots the bin-wise accuracy against the oracle which has perfectly balanced bin-wise prediction accuracy and confidence.

B. Performance Under the GCoSOD Setting

1) Quantitative Comparison: Tab. I shows the performances of the SOTA CoSOD and SOD methods on the proposed CoCA-Common and CoCA-Zero datasets. Most of the methods achieve mIoU of less than 50% on the CoCA-Zero dataset and less than 60% on the CoCA-Common dataset. Without inter-image relationship modeling, the SOD methods tend to find all salient objects rather than the co-salient ones, leading to serious over-segmentation and poor mIoU and MAE.
scores. However, it is worth noting that SOD models are not trailing far behind some of their CoSOD counterparts. For example, EGNet [5] and TEP [6] achieve similar performance to ICNet [20], which has a strong reliance on saliency priors. With the extra co-saliency modeling, CoSOD models achieve relatively better performance than SOD models on the two new CoSOD testing datasets. Benefitting from its small inference size, GCAGC [2] achieves the best performance among existing techniques. Instead of feeding the entire group of images to the model, GCAGC [2] randomly selects five images from the
group for co-salient map generation. Images within the small group (five images in GCAGC [2]) are more likely to have a high variance in terms of the appearance and the salient object category, enabling the model to predict the absence of a co-salient object. Taking the entire group of images as input, our method significantly improved over the baselines (CADC [3], ICNet [20] and DCFM [4]) and outperforms all existing CoSOD and SOD methods on both CoCA-Common and CoCA-Zero datasets.

2) Qualitative Comparison: Fig. 4 shows the predictions of existing CoSOD models and GCT-CADC on the CoCA-Common dataset. It illustrates that existing CoSOD models are not robust to the introduction of “noisy images” to a group of images sharing a common salient object, and they are prone to mistakenly predicting salient objects from the secondary groups as the common salient object. GCAGC [2], though outperforming other CoSOD models on the “noisy images”, tends to predict the absence of co-salient objects regardless of their true existence as a result of its small group size during inference. On the contrary, GCT-CADC is robust to both intra-image and inter-image noisy salient objects while accurately segmenting the common salient objects if present.

C. Calibration Performance Evaluation

The proposed GCT also improves the baseline model in terms of the model calibration degree. Fig. 5 illustrates the reliability diagram of CADC and GCT-CADC (Ours). It can be observed that the prediction-to-confidence ratio of GCT-CADC is closer to 1 (Oracle) than the baseline, indicating a higher calibration degree. Further, Tab. II shows the expected calibration error of the existing CoSOD models and our models (GCT-CADC, GCT-DCFM, GCT-ICNet). Each of our models has significantly improvement over the baseline models under both the conventional and generalised CoSOD settings. Furthermore, GCT-CADC achieves the lowest ECE on four out of five testing datasets.

D. Ablation Study

1) Effectiveness of Generalised CoSOD Training: In Tab. I, we observe significantly improved performance compared with the baseline models [3]. Take GCT-CADC vs. CADC as an example, it enhances the mIoU scores by 34.2% and 109.7%, and reduces the MAE by 48.6% and 90.1% on the CoCA-Common and CoCA-Zero datasets respectively. This is attributed to the successful acknowledgement of the absence of common salient object in the “noisy images”. As a result, GCT-CADC is able to make a full-negative predictions for images where co-salient objects are absent whereas the baseline model tends to output false-positive predictions in “noisy images”. As shown in Fig. 6, the baseline model (“B”) falsely identifies the intra-image and inter-image secondary salient objects as co-salient objects. After incorporating the proposed Generalised CoSOD Training (“B+G”), the false-positive predictions e.g., flower (4th column), helicopter (6th column), lounge (9th column), etc., can be mostly avoided.

2) Quality of Uncertainty Estimation: The random sampling inherent in GCT enables our model to generate high-quality uncertainty maps. As shown in Fig. 6, the uncertainty maps are able to highlight false-positive predictions on secondary salient objects. We further apply a dynamic threshold to the uncertainty map to find potential false-positive co-saliency predictions. Specifically, we compute a bias matrix, measuring the pixel-wise distance between the uncertainty map and its instance-wise mean as \[ \phi = u(x_{ki}) - \mu_{ki} A \], where \( A \) is an all-one matrix, \( \mu_{ki} \) is a single-image uncertainty mean. We remove those unreliable predictions which spatially correspond to positive values in the bias matrix. It can be seen that removing those uncertain predictions further improves the quality of the predictions.

Different from existing uncertainty estimation techniques that usually focus on object boundaries [51]–[53], our method estimates the uncertainties at an instance level. Fig. 7 compares the uncertainty estimation qualities of Ensemble [52], Dropout [51], and Energy-Based Model (EBM) [53] with ours. It can be observed that the compared methods only highlight the boundary of an incorrectly detected non-co-salient object in the noisy image (1st column). Similarly, for the secondary salient objects that are incorrectly identified as a co-salient object, they also only identify the boundary as uncertain (2nd - 5th columns). On the contrary, our method can acknowledge the mis-detection
Fig. 5. Expected Calibration Error of 6 existing CoSOD models [3], [4], [9], [15], [20] and our method (GCT-CADC) on both the conventional CoSOD testing datasets (CoCA [9], CoSOD3k [1], CoSal2015 [26]) and the generalised CoSOD testing datasets (CoCA-Common and CoCA-Zero).

| Method   | CoCA [9] | CoSOD3k [1] | CoSal2015 [26] |
|----------|----------|-------------|-----------------|
|          | F ↑ E ↑ S ↑ M ↓ | F ↑ E ↑ S ↑ M ↓ | F ↑ E ↑ S ↑ M ↓ |
| CADC [3] | 0.548 0.743 0.680 0.133 | 0.759 0.840 0.801 0.096 | 0.862 0.906 0.866 0.064 |
| GCT-CADC | 0.542 0.740 0.683 0.122 | 0.756 0.837 0.798 0.093 | 0.852 0.896 0.853 0.079 |
Fig. 6. Qualitative results of baseline model (“B”), Generalised CoSOD Training (“B+G”) and Uncertainty-based Revision (“B+G+U”) on the “Pocket Watch” group of the CoCA-Common dataset. UC is the uncertainty map. After removing the potential false-positive predictions with high uncertainties, the predictions focus more primarily on the common salient object.

Fig. 7. Comparison of qualitative results of uncertainty maps estimated with Ensemble, Dropout, Energy Based Model (EBM), and Ours. The header indicates the co-salient object of the group, from which the sample is chosen. High uncertainties are represented in red and low uncertainties in blue in the uncertainty maps.

in its entirety as uncertainty in both situations. For example, the heart in the “pine cone” group (1st column); the chilli pepper in the “eggplant” group (3rd column); the pen in the “calculator” group (5th column). Following [54], we adopt ECE to quantitatively evaluate the estimated uncertainty of our GCT and the compared methods. As shown in Tab. II, our GCT achieves higher ECE scores than the compared methods in both CoSOD and GCoSOD settings, indicating a better reliability of our estimated uncertainty map.

E. Discussion

1) Performance for Conventional CoSOD Datasets: We further evaluate the proposed “Generalised CoSOD Training” strategy on conventional CoSOD testing datasets, and show performance in Tab. III using “CADC” as the baseline model. It shows that GCT, while significantly enhancing the performance of the baseline model in terms of handling the “noisy images” in the GCoSOD setting and model calibration degree in both conventional and generalised CoSOD settings, still performs very close to the baseline model under the conventional setting. We also observe improved MAE on
Fig. 8. Example images included in the “apple” group of CoSal2015 [26] (left) and CoSOD3k [1] (right). Intact red apple makes up the majority of the “apple” group of CoSal2015. This renders our model, which is aware of potential noisy salient objects at inter-image level, to discriminate the intact red apple against its variants including chewed red apple, grayscale apple, peeled apple, and green apple. On the other hand, “Apple” group of CoSOD3k [1] contains primarily intact green apple without variants and our model can also acknowledge the universal presence of co-salient object in each individual image.

### TABLE IV
CoCA-Common benchmark with different primary ratios.

| Method    | 20 - 40% mIoU ↑ M ↓ | 40 - 60% mIoU ↑ M ↓ | 60 - 80% mIoU ↑ M ↓ |
|-----------|---------------------|---------------------|---------------------|
| EGNet [5] | 44.85 0.215         | 49.13 0.203         | 53.05 0.194         |
| TEP [6]   | 44.54 0.222         | 49.00 0.209         | 53.89 0.200         |
| ICNet [20]| 46.17 0.208         | 52.06 0.188         | 57.52 0.171         |
| GCAGC [2] | 63.21 0.066         | 61.87 0.080         | 61.42 0.087         |
| GICD [9]  | 48.74 0.156         | 53.49 0.152         | 58.10 0.141         |
| GCoNet [15]| 52.65 0.084        | 56.32 0.096         | 60.89 0.101         |
| CADC [3]  | 48.87 0.153         | 54.88 0.132         | 59.20 0.131         |
| GCT-CADC  | 74.75 0.058         | 72.06 0.078         | 69.01 0.086         |

### TABLE V
ECE of our GCT and the compared uncertainty estimation methods including Dropout [51], Ensemble [52], and EBM [53].

| Method    | CoCA | CoSOD3k | CoSal2015 | CoCA-Common | CoCA-Zero |
|-----------|------|---------|-----------|-------------|-----------|
| Dropout   | 0.087| 0.057   | 0.049     | 0.105       | 0.144     |
| Ensemble  | 0.060| 0.042   | 0.025     | 0.068       | 0.095     |
| EBM       | 0.086| 0.051   | 0.043     | 0.096       | 0.121     |
| GCT-CADC  | **0.058** | **0.036** | **0.016** | **0.022**  | **0.007** |

CoCA [9] and CoSOD3k [1]. The performance drop can be attributed to the awareness of potential absence of common salient object in the “noisy images”. Conventional models require a co-salient object to be present in every image, ideally matching the conventional CoSOD datasets.

GCT training allows a model to predict the absence of a co-salient objects in images, which does not occur in conventional CoSOD testing datasets. In this case, GCT-CADC allows discrimination of a common salient object from its different variants, leading to false negative predictions. Fig. 8 illustrates the false negative predictions of the proposed GCT on the apple group of the CoSal2015 dataset [26]. The awareness of a potential absence of a co-salient object in individual images of the group makes our model classify the target objects in a more granular scale. Specifically, the intact red apple makes up the majority of the apple group in CoSal2015 [26], resulting in our model discriminating against in various apple variants including chewed apple, sliced apple, green apple, peeled apple, and grayscale apple. On the other hand, the apple group in CoSOD3K [1] dataset contains intact green apples with consistent appearances and our model successfully detect them as the co-salient object.

2) Performance w.r.t. Primary Group Ratios:: We divide the CoCA-Common dataset into three different groups with primary group ratio in the range of [0.2, 0.4), [0.4, 0.6) and [0.6, 0.8] respectively, and show the performance of existing CoSOD and SOD models on the above individual sets in Tab. IV. We observe decreased performance with the smaller primary group ratio for existing CoSOD and SOD models. This is due to their implicit assumption that a co-salient object can be found in each individual image. On the contrary, our method acknowledges the absence of co-salient object, leading to better performance in all ranges of primary group ratio. Tab. IV reveals a converging trend between our method and the baseline model as the primary ratio increases. This is expected as GCT-CADC has better detection of the absence of a co-salient object in “noisy images” than its baseline. Their performance is similar when the common salient object is present in each individual image of the same group.

V. Conclusion
We propose the Generalised Co-Salient Object Detection setting where “noisy images” that do not share the common salient object can be present in various degrees. We further design a random sampling based Generalised Training strategy,
using Diverse Sampling Self-Supervised Learning to distill awareness of the potential absence of the common salient object into existing CoSOD models. In addition, the random sampling process inherent in the GCT allows the model to produce high-quality uncertainty maps that highlight the potential false positive predictions at instance level. Our method is compatible with existing CoSOD models and significantly improves existing CoSOD models in terms of: (1) performance under the GCoSOD setting where “noisy images” with no common salient object visible can be present in various degrees; and, (2) the model calibration degree under the standard CoSOD setting, preventing over-confident predictions.

REFERENCES

[1] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen, “Re-thinking co-salient object detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2021.
[2] K. Zhang, T. Li, S. Shen, B. Liu, J. Chen, and Q. Liu, “Adaptive graph convolutional network with attention graph clustering for co-saliency detection,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9050–9059.
[3] N. Zhang, J. Han, N. Liu, and L. Shao, “Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection,” in IEEE International Conference on Computer Vision (ICCV), 2021, pp. 4167–4176.
[4] S. Yu, J. Xiao, B. Zhang, and E. G. Lim, “Democracy does matter: Comprehensive feature mining for co-salient object detection,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 979–988.
[5] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, “Egnet: Edge guidance network for salient object detection,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 8779–8788.
[6] J. Zhang, J. Xie, N. Barnes, and P. Li, “Learning generative vision transformer with energy-based latent space for saliency prediction,” in Conference on Neural Information Processing Systems (NeurIPS), vol. 34, 2021.
[7] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, “Interactive two-stream decoder for accurate and fast saliency detection,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9141–9150.
[8] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in International Conference on Machine Learning (ICML), 2017, pp. 1321–1330.
[9] Z. Zhang, W. Jin, J. Xu, and M.-M. Cheng, “Gradient-induced co-saliency detection,” in European Conference on Computer Vision (ECCV), Springer, 2020, pp. 455–472.
[10] 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 (TIP), vol. 23, no. 9, pp. 4175–4186, 2014.
[11] H. Li and K. N. Ngan, “A co-saliency model of image pairs,” IEEE Transactions on Image Processing (TIP), vol. 20, no. 12, pp. 3365–3375, 2011.
[12] 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, 2014.
[13] H. Fu, X. Cao, and Z. Tu, “Cluster-based co-saliency detection,” IEEE Transactions on Image Processing (TIP), vol. 22, no. 10, pp. 3766–3778, 2013.
[14] D. Zhang, D. Meng, and J. Han, “Co-saliency detection via a self-paced multiple-instance learning framework,” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 39, no. 5, pp. 865–878, 2016.
[15] Q. Fan, D.-P. Fan, H. Fu, C.-K. Tang, L. Shao, and Y.-W. Tai, “Group collaborative learning for co-salient object detection,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12798–12798.
[16] B. Li, Z. Sun, L. Tang, Y. Sun, and J. Shi, “Detecting robust co-saliency with recurrent co-attention neural network,” in International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 818–825.
[17] K. Zhang, T. Li, B. Liu, and Q. Liu, “Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3095–3104.
[18] Q. Zhang, R. Cong, J. Hou, C. Li, and Y. Zhao, “Coadnet: Collaborative aggregation-and-distribution networks for co-salient object detection,” in Conference on Neural Information Processing Systems (NeurIPS), vol. 33, 2020.
[19] K. Zhang, M. Dong, B. Liu, X.-T. Yuan, and Q. Liu, “Deepgc: Co-saliency detection via semantic-aware contrast gromov-wasserstein distance,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13 703–13 712.
[20] W.-D. Jin, J. Xu, M.-M. Cheng, Y. Zhang, and W. Guo, “Icenet: Intra-saliency correlation network for co-saliency detection,” in Conference on Neural Information Processing Systems (NeurIPS), vol. 33, 2020.
[21] L. Wei, S. Zhao, O. E. F. Bourahla, X. Li, F. Yu, and Y. Zhuang, “Deep group-wise fully convolutional network for co-saliency detection with graph propagation,” IEEE Transactions on Image Processing (TIP), vol. 28, no. 10, pp. 5052–5063, 2019.
[22] B. Jiang, X. Jiang, J. Tang, and B. Luo, “Co-saliency detection via a general optimization model and adaptive graph learning,” IEEE Transactions on Multimedia (TMM), vol. 23, pp. 3193–3202, 2020.
[23] B. Jiang, X. Jiang, A. Zhou, J. Tang, and B. Luo, “A unified multiple graph learning and convolutional network model for co-saliency estimation,” in ACM International Conference on Multimedia (MM), 2019, pp. 1375–1382.
[24] J. Winn, A. Criminisi, and T. Minka, “Object categorization by learned universal visual dictionary,” in IEEE Conference on Computer Vision (ICCV), vol. 2, IEEE, 2005, pp. 1800–1807.
[25] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, “iiscog: Interactive co-segmentation with intelligent scribble guidance,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 3169–3176.
[26] 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 (IJCV), vol. 120, no. 2, pp. 215–232, 2016.
[27] T. Miyato, S. Maeda, S. Ishii, and M. Koyama, “Virtual adversarial training: A regularization method for supervised and semi-supervised learning,” IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), pp. 1–1, 2018.
[28] T. Miyato, S. Maeda, M. Koyama, K. Nakae, and S. Ishii, “Distributional smoothing with virtual adversarial training,” in International Conference on Learning Representations (ICLR), 2016.
[29] J. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” CoRR, abs/1412.6572, 2015.
[30] A. Kumar, S. Sarawagi, and U. Jain, “Trainable calibration measures for neural networks from kernel mean embeddings,” in International Conference on Machine Learning (ICML), 2018, pp. 2805–2814.
[31] D. Tu, F. E. T. Jaynes, “Information theory and statistical mechanics,” Physical Review, vol. 106, no. 4, p. 620, 1957.
[32] Z. Zhang, W. Jin, J. Xu, and M.-M. Cheng, “Gradient-induced co-saliency detection,” in European Conference on Computer Vision (ECCV), Springer, 2020, pp. 12798–12798.
[39] J. Liu, J. Zhang, and N. Barnes, “Modeling aleatoric uncertainty for camouflaged object detection,” in Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1445–1454.

[40] P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. G. Wilson, “What are bayesian neural network posteriors really like?” in International Conference on Machine Learning (ICML). PMLR, 2021, pp. 4629–4640.

[41] S. W. Ober and L. Aitchison, “Global inducing point variational posteriors for bayesian neural networks and deep gaussian processes,” in International Conference on Machine Learning (ICML). PMLR, 2021, pp. 8248–8259.

[42] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, “Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning,” in International Conference on Machine Learning (ICML). PMLR, 2018, pp. 1184–1193.

[43] Y. Kwon, J.-H. Won, B. J. Kim, and M. C. Paik, “Uncertainty quantification using bayesian neural networks in classification: Application to biomedical image segmentation,” Computational Statistics & Data Analysis, vol. 142, p. 106816, 2020.

[44] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, “Structure-measure: A new way to evaluate foreground maps,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4548–4557.

[45] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, “Enhanced-alignment measure for binary foreground map evaluation,” in International Joint Conference on Artificial Intelligence (IJCAI), 2018, pp. 698–704.

[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in European Conference on Computer Vision (ECCV). Springer, 2014, pp. 740–755.

[47] M. H. Degroot and S. E. Fienberg, “The comparison and evaluation of forecasters.” The Statistician, vol. 32, pp. 12–22, 1983.

[48] A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” in International Conference on Machine Learning (ICML), 2005, p. 625–632.

[49] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in International Conference on Machine Learning (ICML). PMLR, 2017, pp. 1321–1330.

[50] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in International Conference on Machine Learning (ICML). PMLR, 2016, pp. 1050–1059.

[51] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in Conference on Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[52] J. Zhang, J. Xie, N. Barnes, and P. Li, “Learning generative vision transformer with energy-based latent space for saliency prediction,” in Conference on Neural Information Processing Systems (NeurIPS), vol. 34, 2021, pp. 15448–15463.

[53] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, “Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift,” in Conference on Neural Information Processing Systems (NeurIPS), vol. 32, 2019.