Unsupervised Salient Instance Detection

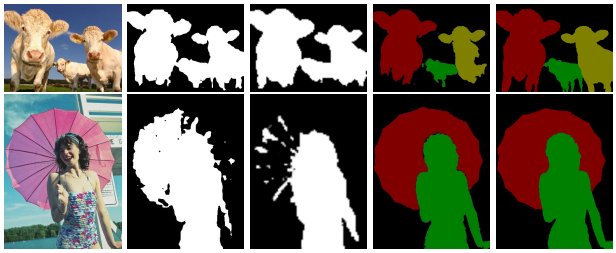
Xin Tian[1] Ke Xu[2,†] Rynson Lau[2,†]
[1]Huawei Technologies [2]City University of Hong Kong
{xin.tian.831, kkangwing}@gmail.com, Rynson.Lau@cityu.edu.hk
[†] joint corresponding authors

Abstract

The significant amount of manual efforts in annotating pixel-level labels has triggered the advancement of unsupervised saliency learning. However, without supervision signals, state-of-the-art methods can only infer region-level saliency. In this paper, we propose to explore the unsupervised salient instance detection (USID) problem, for a more fine-grained visual understanding. Our key observation is that self-supervised transformer features may exhibit local similarities as well as different levels of contrast to other regions, which provide informative cues to identify salient instances. Hence, we propose SCoCo, a novel network that models saliency coherence and contrast for USID. SCoCo includes two novel modules: (1) a global background adaptation (GBA) module with a scene-level contrastive loss to extract salient regions from the scene by searching the adaptive "saliency threshold" in the self-supervised transformer features, and (2) a locality-aware similarity (LAS) module with an instance-level contrastive loss to group salient regions into instances by modeling the in-region saliency coherence and cross-region saliency contrasts. Extensive experiments show that SCoCo outperforms state-of-the-art weakly-supervised SID methods and carefully designed unsupervised baselines, and has comparable performances to fully-supervised SID methods.

1. Introduction

Salient Instance Detection (SID) aims to detect instance-level salient objects. It can provide a fine-grained visual understanding of the scene, and is therefore able to facilitate a wide range of applications, *e.g.*, image captioning [21], scene text recognition [20], medical image analysis [19], and others [1, 4, 43, 48, 91]. However, existing SID methods either require pixel-level labels [25, 74] or image-level labels [62] as supervisions, which are expensive to prepare. Despite the advancements in unsupervised saliency detection [32, 44, 57, 72, 79, 81, 87, 88], none of the existing methods have demonstrated the capability to learn instance-level saliency, which essentially requires fine-grained differentiation not only between salient and non-salient regions but also between different salient regions (Figure 1). In this paper, we propose to study this unexplored task, *i.e.*, unsupervised salient instance detection (USID).



(a) Image (b) FOUND [57] (c) SelfMask [55] (d) Ours (e) GT

Figure 1. Existing unsupervised salient object detection methods, FOUND [57] (b) and SelfMask [55] (c), cannot differentiate salient instances (rows 1 and 2), but tend to detect strong semantic instances, *e.g.*, the woman in row 2. In contrast, our method (d) correctly detects all salient instances by modeling saliency coherence and different levels of saliency contrast.

Our key observation of this problem is built upon the advanced self-supervised transformer representations. We observe that their features exhibit **saliency coherence** (high similarities) within local regions and may show varying degrees of **saliency contrast** (lower similarities and dissimilarities) between different regions. In Figure 2, we visualize global saliency information within the self-supervised transformer features (b-d), and the saliency coherence and contrast between three feature pixels and the rest of the image (e-g), where each of these pixels is randomly selected from a salient instance. While the transformer features themselves may exhibit some degree of global saliency contrast (b-d), the computed pixel-based correlations show that each selected pixel tends to have high saliency coherence with pixels of the same instance but different levels of saliency contrast to pixels of other instances and the background (e-g). This observation inspires us to mine regional saliency coherence and contrast information for USID.

In this paper, we propose a novel approach, called *SCoCo*, to model Saliency Coherence and Contrast information from deep self-supervised features to address the USID problem. Specifically, SCoCo first models scene-level saliency contrast to determine the "saliency thresh-

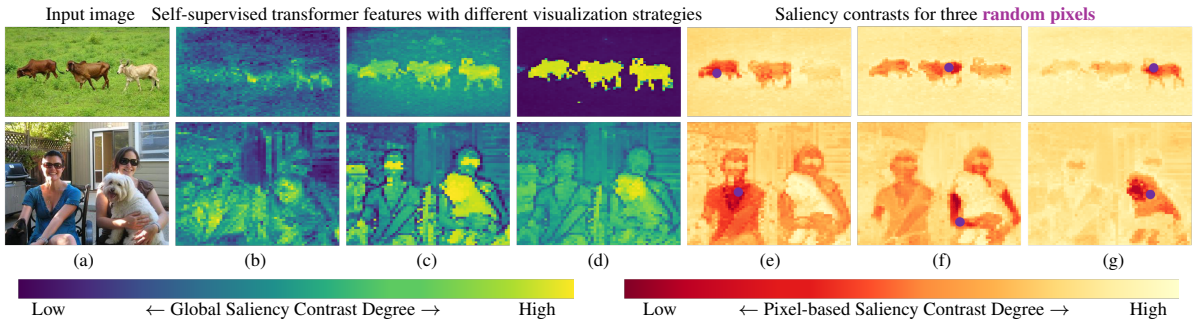


Figure 2. Given an input image (a), we first visualize its corresponding self-supervised transformer features [8] (b), the pixel correlations to the [CLS] token (c), and the pixel correlations over the global scene (d). We then randomly sample three pixels (**purple dots**), each for one instance, and visualize the saliency contrast of the the transformer features between the sampled pixel and all other pixels of the image by computing feature point correlations (e-g). We observe that while the self-supervised transformer features may show the global saliency contrast to some degree, pixels from the same instance tend to be more coherent, while pixels from different regions tend to have different levels of contrast. This inspires us to mine such regional saliency coherence and contrast for USID.

old," which differentiates salient regions with respect to diverse background contents. We implement this via a novel global background adaptation (GBA) module with a scene-level contrastive loss, which learns to iteratively pull the saliency regions out of the complex background. SCoCo then models the within-region saliency coherence and cross-region saliency contrast information for grouping pixels into salient instances. We implement this by proposing a novel locality-aware similarity (LAS) module with an instance-level contrastive loss, which initially focuses on modeling the saliency coherence within individual regions and then shifts to model different levels of saliency contrast between different regions.

To summarize, this paper has three main contributions:

- To our knowledge, we propose the first USID approach, named SCoCo, to learn to detect salient instances without any annotated labels, by modeling the intra-region coherence and inter-region contrast information from self-supervised transformer features.
- SCoCo has two novel modules: a GBA module with a scene-level contrastive loss to find the saliency thresholds of different image contents, and a LAS module with an instance-level contrastive loss to detect salient instances.
- Extensive experiments show that SCoCo outperforms existing weakly-supervised SID methods and carefully designed unsupervised baselines, and has comparable performances to fully-supervised SID methods.

2. Related Works

2.1. Salient Instance Detection (SID)

SID aims to individualize visually prominent objects in the input image. Existing deep SID methods can be classified into two categories based on their learning labels: fully-supervised and weakly-supervised SID methods.

Fully-supervised SID methods rely on costly pixel-wise annotated masks for training supervision. Li *et al*. [25] propose to apply MAP [80] on instance proposals, which are generated based on binary saliency and contour maps, to select and segment salient instances. Fan *et al*. [14] devise a single-stage salient instance detection method, with a RoIMasking layer that utilizes both local and surrounding contexts of each instance for detection. Pei *et al*. [45] propose a spectral clustering method to obtain salient instances by exploring pixel masks and subitizing information of salient objects, and later a transformer-based method [46] to detect salient instances in a single stage without NMS for post-processing. Liu *et al*. [37] propose an interleaved execution strategy to incorporate global context and object contour information jointly to detect salient instances. RGB-D information has also been explored in CalibNet *et al*. [47], with a dual-branch cross-modal calibration method.

Weakly-supervised SID methods typically use bounding boxes and image-level labels, *e.g*., classes and subitizing as supervisions. Zhang *et al*. [80] propose a Maximum A Posteriori (MAP)-based subset optimization formulation to select a compact set of salient instances from the redundant box-level proposals, followed by CRF to help segment pixel-wise instances from box-level predictions. Tian *et al*. [61, 62] propose a triple-branch network to exploit salient boundaries, centroids, and regions jointly from class and subitizing labels for salient instance detection.

Despite their success, existing SID methods require human efforts to prepare different types of labels, which are expensive. In contrast, we propose to model the saliency coherence and contrast information based on self-supervised transformer features, for USID.

2.2. Salient Object Detection (SOD)

SOD is an important task to scene understanding with a lot of methods proposed. Deep SOD methods can be divided into three groups w.r.t. the supervision signals: fully-supervised, weakly-supervised, and unsupervised.

Fully-supervised SOD methods mainly incorporate four types of deep techniques to learn salient object represen-

tations from pixel-level annotated masks. First, deep feature fusion [34, 39, 68, 73, 83] is used to aggregate multi-level context information including low-level stimulus and high-level semantics for detecting salient objects. Second, attention mechanisms [30, 35, 36, 58, 84, 85] are used to reweight multi-scale features and enhance context learning to help the model focus on the salient regions and suppress noise of the background regions. Third, boundary enhancement [9, 33, 52, 86] is often applied to improve the localization and segmentation results by penalizing the errors of boundary-surrounding pixels. Last, recurrent mechanisms [22, 29, 60, 66, 84] may also be reused to iterate an identical network to predict and refine the saliency regions in a coarse-to-fine manner.

Weakly-supervised SOD methods have explored class labels, image captions, scribbles, bounding boxes, and points as supervisions. Wang *et al*. [67] propose the first weakly-supervised SOD method, which leverages image-level class labels to localize salient regions and then refines those predicted regions with CRF. Following [67], Li *et al*. [26] propose a pseudo saliency annotation updating scheme to refine the spatial coherence of predicted salient regions. Piao *et al*. [50] propose the multi-filter directive network to distill clean saliency maps from multiple class-based pseudo labels. Zeng *et al*. [78] devise a multi-source weak supervision framework to jointly utilize category labels, captions, and a set of unlabelled images for training. Based on scribble annotations, WSSA [82] and AGGM [77] are proposed to detect salient objects by exploring local consistency and structure information of objects. Liu *et al*. [40] propose to use bounding box supervisions to help address the over/under detection problems. Gao *et al*. [15] design an adaptive masked flood filling algorithm to detect salient objects from point-wise labels.

Unsupervised SOD methods can be classified into four categories according to the features and technologies used. First, conventional methods are typically based on hand-crafted features [2, 5, 38, 49, 76, 92], *e.g*., local and global contrast [10, 38, 49], boundary priors [27, 76, 92], and spatial frequency [2], for detecting salient objects. These methods may not handle complex scenes as their low-level features have limited representation capacities. Second, a series of methods [32, 44, 72, 79, 81, 87, 88] try to distill clean labels from pseudo labels, which are produced by some of the aforementioned conventional methods, for learning SOD models. Specifically, they use distillation strategies like iterative refinement [44], causal debiasing [32], and mining textures around boundaries [89] to reduce the noise coming from the pseudo labels. Third, instead of learning from noisy pseudo labels, Yan *et al*. [75] propose a domain adaptive SOD method that learns from synthetic but clean saliency data. Fourth, FOUND [57] and SelfMask [55] also detect salient objects based on self-supervised deep features. However, our SCoCo differs from them in both tasks and proposed techniques. First, they cannot predict instance-level saliency information, as shown in row 1 of Figure 1(b,c). Second, they only model the binary contrast between salient and non-salient regions; they are unable to suppress non-salient regions/objects with uncertain contrast degrees. For example, in row 2 of Figure 1(c,d), FOUND and SelfMask focus mainly on the most prominent woman, neglecting the umbrella, as the woman has much stronger semantics than the umbrella. In contrast, our SCoCo (Figure 1(d)) can correctly individualize salient instances by modeling intra-region saliency coherence and different levels of saliency contrast between regions.

In summary, unlike the above SOD methods, which do not consider instance-level object saliency, our work aims to detect salient instances but without the need to use annotations for supervision.

2.3. Unsupervised Object Detection (UOD)

UOD aims to localize objects without using any human annotations. Early methods [11, 63–65] explore inter-image similarities (*e.g*., clustering, ranking) to discover objects for an image collection. Recently, several methods learn to detect objects based on self-supervised deep features. LOST [56] and TokenCut [71] propose to segment one object per image based on the pixel graph of self-supervised deep features. FreeSOLO [69] and CutLER [70] learn to detect multiple objects by self-retraining [69] and MaskCut [70].

While these methods tend to detect any objects (salient, non-salient), our method learns to identify salient instances.

3. Methodology

Learning an unsupervised salient instance detector is challenging, as there are no explicit labels/knowledge to teach the detector to identify salient regions at the instance level. In this work, we observe some interesting cues in the self-supervised transformer features – they exhibit saliency coherence within local regions but saliency contrast of diverse degrees between different regions. If a detector can model such correlations well, it may be able to identify salient regions and group them into separate salient instances. Here, we propose the SCoCo method to mine regional saliency coherence and contrast information for USID.

Figure 3 shows the overview of our SCoCo. Given an input image, we first apply a self-supervised transformer (with frozen weights) to generate the corresponding deep features. We then feed these deep features separately to a novel global background adaptation (GBA) module (Section 3.1) and a novel locality-aware similarity (LAS) module (Section 3.2). The GBA module with its scene-level contrastive loss aims to find the suitable 'saliency thresholds' to distinguish salient regions from diverse background contents to produce a scene-level saliency map. Meanwhile,

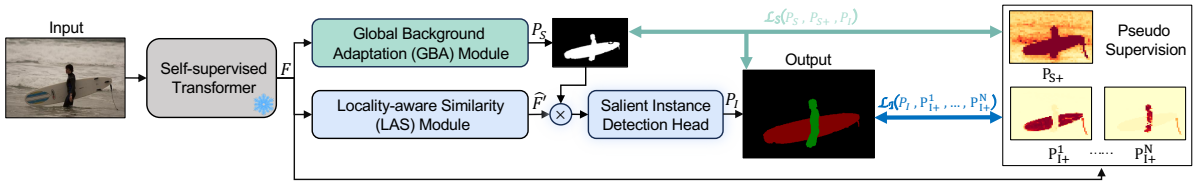


Figure 3. The overview of our unsupervised pipeline. Given an image, we first obtain its features F from a self-supervised transformer (whose weights are frozen). We then use F to produce pseudo scene- and instance-level supervisions $\mathrm{P_{S+}}$ and $\mathrm{P}^n_{\mathrm{I+}}$ ($1 \leq n \leq N$) by computing feature similarities, and feed F to our novel GBA and LAS modules. The GBA module learns with a scene-level contrastive loss $\mathcal{L}_{\mathcal{S}}$ to predict a saliency map $\mathrm{P_S}$, while the LAS module learns with an instance-level contrastive loss $\mathcal{L}_{\mathcal{I}}$ to model saliency coherence and contrast features. The output features of both GBA and LAS modules are combined in the salient instance detection head to produce the final salient instance map $\mathrm{P_I}$.

the LAS module works with an instance-level contrastive loss to learn intra-region saliency coherence and inter-region saliency contrast features. These features are then fed into a salient instance detection head to produce salient instances, with the guidance of the scene-level saliency map produced by the GBA module.

3.1. Global Background Adaptation (GBA) Module

The goal of the GBA module is to adaptively dig saliency information from diverse background contents. As Figure 2(b-d) shows, it is challenging to find a fixed saliency threshold to select salient regions from different scenes with diverse complexities. To address this issue, we propose the GBA module, which has learnable saliency threshold functions working in a bifurcation-and-combination manner, to detect the fine-grained adaptive saliency thresholds. Specifically, as shown in Figure 4, the GBA module is designed to have two bifurcated branches at the first, with each of them equipped with its own learnable saliency threshold function, to handle the potential intra-scene saliency-background ambiguities separately. The two bifurcated branches are then combined with another saliency threshold-based branch to produce the scene-level saliency map.

Our GBA module differs from previous unsupervised salient object detection methods [55, 57] in the way of detecting saliency thresholds. Existing USOD methods typically rely on stacking convolutional operations to detect salient objects. However, such operations (even with dynamic convolution kernels [24, 59, 90]) are content-agnostic as convolution filters are shared across all the feature pixels. As a result, these methods often fail to detect the salient regions buried in diverse background contents.

Network Structure. Given features $F \in \mathbb{R}^{H\times W\times C}$, we first process it with two parallel branches, each consisting of a group of 1×1 and 3×3 conv operators, and a learnable saliency threshold function $f(x)$. We define $f(x)$ as:

$$f(x) = (p_1 - p_2)x \cdot \sigma(\theta(p_1 - p_2)x) + p_2 x, \qquad (1)$$

where $f(x)$ is a smooth version of Swish rectifying function [42, 53], p_1 and p_2 are learnable parameters that can jointly determine the upper and lower bounds of the function, σ is the Sigmoid function, and the learnable parameter θ controls the "selection degree" of salient objects with respect to the background. Each θ is learned using two fully connected (FC) layers upon a global image vector generated by the global average pooling (GAP) layer. We then use another branch to combine the former two branches, with another set of convolutional operations for feature transformation and learnable saliency threshold function to unify the separately modeled saliency thresholds, to produce the scene-level saliency map.

Scene-level Contrastive Loss. To boost the saliency learning of the GBA module, we introduce a scene-level contrastive loss $\mathcal{L}_{\mathcal{S}}$. Since the image background is expected to have low saliency degrees, we first derive the scene-level pseudo saliency supervision $\mathrm{P_{S+}}$ through inverting the identified background regions with low saliency degrees. However, we note that such pseudo-saliency maps may contain a significant number of noisy pixels, particularly those false positive background pixels with medium saliency degrees, that may hinder learning efficacy. To address this problem, we leverage the pixels with high saliency degrees in our instance-level saliency predictions $\mathrm{P_I}$ (introduced in Section 3.2) to help classify those uncertain pixels into salient and non-salient ones. To this end, we take $\mathrm{P_I}$ as positive samples and utilize scene-level saliency in different images as negative samples. $\mathcal{L}_{\mathcal{S}}$ can then be formulated as:

$$\mathcal{L}_{\mathcal{S}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{(\mathrm{d}(\mathrm{P}^i_{\mathrm{S}},\mathrm{P}^i_{\mathrm{S+}})+\mathrm{d}(\mathrm{P}^i_{\mathrm{S}},\mathrm{P}^i_{\mathrm{I}}))/\tau}}{e^{(\mathrm{d}(\mathrm{P}^i_{\mathrm{S}},\mathrm{P}^i_{\mathrm{S+}})+\mathrm{d}(\mathrm{P}^i_{\mathrm{S}},\mathrm{P}^i_{\mathrm{I}}))/\tau} + \sum_{k=1,k\neq i}^{N} e^{\mathrm{d}(\mathrm{P}^i_{\mathrm{S}},\mathrm{P}^k_{\mathrm{S+}})/\tau}}, \qquad (2)$$

where N is the batch size, $\mathrm{P_S}$ is the scene-level saliency prediction, d measures the mean square error (MSE), and τ is a temperature parameter.

3.2. Locality-aware Similarity (LAS) Module

The LAS module aims to capture regional saliency coherence information from the self-supervised deep features. This information can then be used to help learn different levels of saliency contrast using an instance-level contrastive loss. We note that the latest saliency instance de-

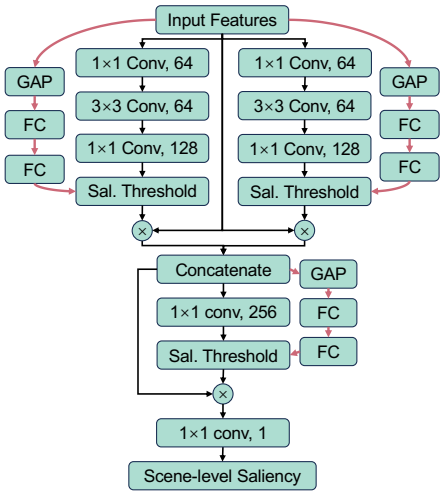


Figure 4. The bifurcation-to-combination learning structure of the global background adaptation (GBA) module. The pink lines indicate the learning flows of learnable saliency threshold functions for extracting salient objects from different background contents.

tection methods [37, 46] typically leverage the object detector head [6, 17] (a combination of convolutions and self-attention blocks) to classify and group salient instance pixels. However, although convolutional filters are good at detecting repeated patterns, they may fail in discerning pixel similarities. The self-attention mechanisms tend to model the interactions of pixel tokens but do not compare region-level saliency contrasts.

On the other hand, our LAS module first models intra-region feature coherence by computing channel-wise correlations with respect to regions surrounding every pixel location, and then feeds regional features to an instance-level contrastive loss to learn regional saliency contrasts for discriminating salient instances. The overview of LAS module is depicted in Figure 5.

Network Structure. Given features $F \in \mathbb{R}^{H\times W\times C}$, we downsample it to $F' \in \mathbb{R}^{H\times W\times C'}$ with a linear layer. We then compute cosine similarity between a positional feature $x \in [1, H] \times [1, W]$ and the surrounding region of x in F', and collect them into a similarity feature tensor $S \in \mathbb{R}^{H\times W\times M\times M\times C'}$. The formulation of S is:

$$S(x, m) = \frac{F'_{(x)} \cdot F'_{(x+m)}}{\|F'_{(x)}\| \cdot \|F'_{(x+m)}\|}, \tag{3}$$

where $m \in [-M + 1, M - 1] \times [-M + 1, M - 1]$, and $M \times M$ is the region size. Eq. 3 aims to encode the local saliency coherence of each pixel x into S. We further apply convolutional filters on S to compact regional features. The spatial size of S is progressively aggregated from $M \times M$ to 1×1 with cascaded 3×3 convolutional filters (without padding) to obtain S' with the same spatial size as the input feature F'. After each convolutional filter, we also insert the batch normalization and rectifying linear unit layers. Finally, we combine F' and S' to produce the final features $\hat{F}'$ with saliency coherence information, which further facilitates the saliency contrast modeling for differentiating salient instances.

Instance-level Contrastive Loss. We propose the instance-level contrastive loss $\mathcal{L}_{\mathcal{I}}$ to derive the saliency contrast information from $\hat{F}'$ for salient instance discrimination, as the binary cross entropy (BCE) loss adopted by existing unsupervised SOD methods [56, 69, 70] cannot distinguish different salient instances.

Given the self-supervised transformer features, we extract the key embeddings F_{key}, and feed the self-affinity matrix of F_{key} to Normalized Cut [54, 71] to find highly coherent salient pixels as pseudo salient instance supervision $\mathrm{P}_{\mathrm{I}+}$ in an iterative manner. We find one salient instance $\mathrm{P}^k_{\mathrm{I}+}$ in each iteration $k \in [1, K]$, and the iteration ends if the $\mathrm{P}^{K+1}_{\mathrm{I}+}$ is located in the low-saliency background. Note that P_I is used in $\mathcal{L}_{\mathcal{S}}$ (Eq. 2) for penalizing pixels with uncertain saliency degrees. Meanwhile, we can measure the spatial discrepancy between predicted instances $\mathrm{P}^L_{\mathrm{I}}$ and $\mathrm{P}^K_{\mathrm{I}+}$ to formulate $\mathcal{L}_{\mathcal{I}}$, where L in the number of instances. To reduce the influence of noise that may be contained in pseudo labels, we propose to sample point vectors from $\hat{F}'$ to compute $\mathcal{L}_{\mathcal{I}}$, where $\hat{F}'$, $\mathrm{P}^L_{\mathrm{I}}$, and $\mathrm{P}^K_{\mathrm{I}+}$ are spatially aligned for the sampling. By sampling a small subset of point vectors, we make the model less likely to be influenced by extreme noise that may be present in the full pseudo label set. Specifically, given $\hat{F}'$, we randomly sample T ($t \in [1, T]$) triplets of point vectors $\hat{F}'^{\,l(t)}_{\mathrm{I}}$, $\hat{F}'^{\,l(t)}_{\mathrm{I}+}$, and $\hat{F}'^{\,l(t)}_{\mathrm{I}-}$ from $\hat{F}'$with respect to each predicted instance $\mathrm{P}^l_{\mathrm{I}}$, its positive supervision $\mathrm{P}^k_{\mathrm{I}+}$ (IoU $(\mathrm{P}^l_{\mathrm{I}}, \mathrm{P}^k_{\mathrm{I}+}) \geq 0.7$), and negative supervision $\mathrm{P}^f_{\mathrm{I}-}$ ($f \neq k, f \in [1, K]$), respectively. $\mathcal{L}_{\mathcal{I}}$ can then be formulated as:

$$\mathcal{L}_{\mathcal{I}} = -\frac{1}{L \times T^2} \sum_{l=1}^{L} \sum_{t=1}^{T} \sum_{t=1}^{T} \log(g(\hat{F}', l, t)), \tag{4}$$

where $g(\cdot)$ is a contrastive estimation function, which is formulated as:

$$g(\hat{F}', l, t) = \frac{e^{sim\langle \hat{F}'^{\,l(t)}_{\mathrm{I}}, \hat{F}'^{\,l(t)}_{\mathrm{I}+} \rangle / \tau}}{e^{sim\langle \hat{F}'^{\,l(t)}_{\mathrm{I}}, \hat{F}'^{\,l(t)}_{\mathrm{I}+} \rangle / \tau} + \sum_{t=1}^{T} e^{sim\langle \hat{F}'^{\,l(t)}_{\mathrm{I}}, \hat{F}'^{\,l(t)}_{\mathrm{I}-} \rangle / \tau}}, \tag{5}$$

where $sim\langle\cdot\rangle$ is the cosine similarity function to compute the distance between two point vectors $\hat{F}'^{\,l(t)}_{\mathrm{I}}$ and $\hat{F}'^{\,l(t)}_{\mathrm{I}+}$, and $l(t)$ represents the t_{th} point vector of the l_{th} instance. Each vector is extracted from $\hat{F}'$ according to its spatial location. Our instance-level contrastive loss (Eq. 4) helps the network learn the amplified saliency contrast of different instances while keeping the saliency coherence within each instance by maintaining their self-similarities.

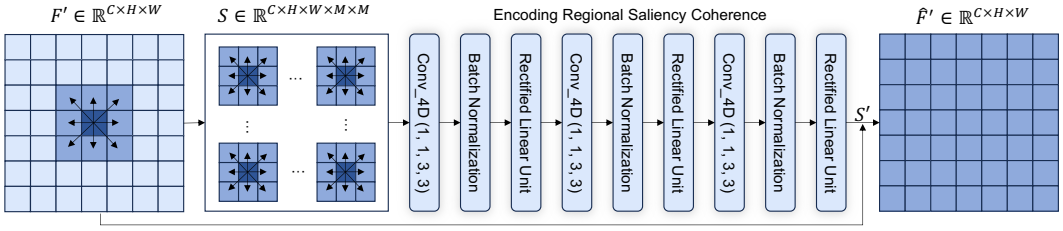


Figure 5. The structure of locality-aware similarity (LAS) module.

4. Experiments

4.1. Experimental Settings

Implementation Details. We use DINO [8] to initialize the self-supervised transformer (*i.e.*, ViT [13]) of our SCoCo network. Note that neither DINO nor our SCoCo uses any labeled data (*e.g.*, class labels) for supervision. The input to both GBA and LAS modules, *i.e.*, F, is the key features of multi-head attention layers in the last feature block of DINO-based ViT. In the GBA module, we use the mean attention value of F as the low saliency degree per image to find its scene-level saliency region P_{S+}. In the LAS module, we sample $T = 32$ triplets of point vectors to formulate the loss $\mathcal{L}_{\mathcal{I}}$. τ in both $\mathcal{L}_{\mathcal{S}}$ and $\mathcal{L}_{\mathcal{I}}$ is set to 0.1. We adopt the head of Mask-RCNN [17] as our salient instance detection head. During training, each input image has a resolution of 1024×1024, and is augmented using large-scale jittering [16]. The batch size N is set to 4 and we use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with step-wise learning rate decay for model optimization. We first train our SCoCo network for 10 epochs using the pseudo supervision derived from DINO. We then update the pseudo labels using SCoCo. Last, we train SCoCo with the updated pseudo labels for another epoch. SCoCo is trained on two 3090ti GPU cards.

Datasets and Evaluation Metrics. We conduct our experiments on three SID datasets, including ILSO-1K [25], ILSO-2K [25], and SIS [46]. The (training/validation/test) sets in the ILSO-1K [25], ILSO-2K [25], and SIS [46] contain (500/200/300), (1000/400/600), and (7030/2100/1170) images, respectively. For each dataset, we only use the unlabeled images of the training set to train our SCoCo network, and we evaluate the performance on the corresponding test set. We adopt the mean Average Precision (mAP) metric for evaluation, following previous SID approaches.

4.2. Baselines

Since we propose the first unsupervised method to detect salient instances, to evaluate our method comprehensively, we design several baselines from three aspects, inspired by related areas, as follows:

- First, we note that existing unsupervised object detection (OD) methods, *i.e.*, FreeSOLO [69], and CutLER [70], detect objects in a saliency-agnostic way. We retrain their models with our pseudo salient instance labels for comparison. Their models are initialized with the model parameters trained on their pseudo labels extracted from large-scale datasets, *e.g.*, COCO [31] and ImageNet [23].
- Second, VitDet [28] is the best-performing method among fully-supervised object detection methods that are based on the ViT [13] backbone. Since we also use ViT as our backbone, we retrain VitDet with our salient instance pseudo labels for comparison.
- Third, we adapt existing USOD methods, *i.e.*, SelfMask [55] and FOUND [57], to detect salient instances. To derive instance-level saliency predictions, we compute IoU between the region-level saliency predictions from USOD methods and the instance predictions from CutlER [70] (a state-of-the-art unsupervised object detection method).

4.3. Main Results

We compare our method to 19 state-of-the-art SID methods, including seven fully-supervised methods: MSRNet-V1 [25], MSRNet-V2 [25], S4Net [14], SCG [37], MDNN [45], RDPNet [74], and OQTR [46]; six weakly-supervised methods: MAP [80], NLDF [41], DeepMask [21], PRM+D [12], IRN [3], and WSID-Net [62]; and six unsupervised baselines: FreeSOLO [69], CutLER [70], VitDet [28] initialized with the classification backbone, VitDet[28] initialized with the DINO-based backbone, SelfMask [55], and FOUND [57].

Quantitative Comparisons. Table 1 shows the quantitative results. First, we can see that our method outperforms all carefully designed unsupervised baselines (the bottom group). This shows that simply retraining unsupervised methods with our pseudo labels and/or filtering out non-salient objects predicted by unsupervised object detectors is not effective in deriving accurate salient instance predictions. Second, we can also see that our method surpasses all existing weakly-supervised methods (middle group), while achieving comparable performances to the fully-supervised methods (first group). Existing weak supervision signals (*i.e.*, object-level pixel masks, bounding boxes, classes, and subitizing labels) cannot provide instance-level supervision, which limits the performance of existing weakly-supervised methods. In contrast, even without supervision, SCoCo can learn to derive instance-level saliency information by modeling saliency coherence and contrast information based on

Table 1. Quantitative comparison with 7 fully-supervised methods, 6 weakly-supervised methods, and 6 unsupervised baselines mentioned in Section 4.2. Column 1 shows the supervision and annotation types of the methods listed in Column 2. Column 3 shows their original tasks (SID: salient instance detection, SOD: salient object detection, OD: object detection, SIS: semantic instance segmentation). Column 4 shows backbone initialization choices, where Cls denotes that their backbones were pre-trained on class labels of ImageNet, and DINO is a self-supervised backbone pretraining method. - denotes missing results as codes/results are not publicly available. For each supervision category, best results are marked in **bold**, while second-best results are underlined.

Supervision type & Annotation type	Method	Original task	Init.	Datasets and Metrics (mAP$_{IoU}$ ↑)					
				ILSO-1K [25]		ILSO-2K [25]		SIS [46]	
				mAP$_{50}$	mAP$_{70}$	mAP$_{50}$	mAP$_{70}$	mAP$_{50}$	mAP$_{70}$
Full Supervison [Images are labeled with instance-level pixel masks.]	MSRNet-V1 [25]	SID	Cls	65.3%	52.1%	-	-	-	-
	MSRNet-V2 [25]	SID	Cls	85.1%	74.7%	78.3%	66.5%	-	-
	S4Net [14]	SID	Cls	82.2%	59.6%	73.1%	52.9%	86.7%	63.6%
	SCG [37]	SID	Cls	88.8%	78.5%	79.8%	68.9%	83.4%	71.2%
	MDNN [45]	SID	Cls	84.9%	67.8%	76.1%	63.6%	84.6%	67.4%
	RDPNet [74]	SID	Cls	88.9%	73.8%	80.7%	67.2%	82.0%	69.4%
	OQTR [46]	SID	Cls	**89.2%**	**81.0%**	**81.4%**	**72.0%**	**88.1%**	**81.7%**
Weak Supervision [Images are labeled with region-level pixel masks*, bounding boxes$^{\triangle}$, classes$^{\circ}$, or subitizing$^{\bowtie}$.]	MAP$^{\triangle}$ [80]	SID	Cls	56.6%	24.8%	51.6%	30.3%	58.4%	22.7%
	NLDF* [41]	SOD	Cls	45.5%	24.5%	43.8%	25.2%	46.4%	24.2%
	DeepMask$^{\triangle}$ [51]	OD	Cls	37.1%	20.5%	35.4%	18.4%	36.4%	21.3%
	PRM+D$^{\circ\bowtie}$ [12]	SIS	Cls	49.6%	31.2%	43.7%	29.8%	52.9%	34.2%
	IRN$^{\circ}$ [3]	SIS	Cls	57.1%	37.4%	50.2%	38.4%	60.1%	40.8%
	WSID-Net$^{\circ\bowtie}$ [62]	SID	Cls	**68.3%**	**51.7%**	**59.4%**	**44.6%**	**68.4%**	**47.1%**
No Supervision [Million-level$^{\clubsuit}$, hundred thousand-level$^{\spadesuit}$, or thousand-level$^{\blacksquare}$ unlabeled images are used for training.]	FreeSOLO$^{\spadesuit}$ [69]	OD	Cls	56.8%	43.7%	50.7%	35.8%	55.4%	43.1%
	FreeSOLO$^{\clubsuit}$ [70]	OD	DINO	53.2%	40.2%	46.9%	32.6%	51.8%	40.3%
	CutLER$^{\clubsuit}$ [70]	OD	Cls	68.6%	54.8%	63.2%	48.3%	66.9%	54.4%
	CutLER$^{\clubsuit}$ [70]	OD	DINO	65.4%	52.1%	59.8%	44.9%	62.4%	51.5%
	VitDet$^{\blacksquare}$ [28]	SIS	Cls	55.6%	43.6%	51.9%	35.8%	54.4%	44.1%
	VitDet$^{\blacksquare}$ [28]	SIS	DINO	55.7%	43.4%	52.4%	37.0%	56.5%	44.8%
	SelfMask$^{\blacksquare}$ [55]	SOD	DINO	65.5%	51.1%	58.9%	44.0%	64.8%	52.1%
	FOUND$^{\blacksquare}$ [57]	SOD	DINO	66.2%	51.3%	61.4%	45.9%	67.1%	53.8%
	Ours$^{\blacksquare}$	SID	DINO	**70.7%**	**58.3%**	**64.8%**	**50.4%**	**71.4%**	**55.7%**

just the self-supervised Transformer features.

Qualitative Comparison. Figure 6 compares the visual results of our method and 10 top performing methods from the three groups of methods in Table 1. We can see that the unsupervised baselines (columns 8 to 11) often fail to detect the complete salient instances and produce messy predictions. For the weakly-supervised methods (columns 5 to 7), MAP, which is trained with bounding box labels, may fail to segment complete objects even with the CRF post-processing. IRN and WSID-Net, which are trained with different class labels, may struggle to determine the boundaries between salient instances of the same class, *e.g.*, the sausages in row 3. Fully-supervised methods (columns 2 to 4) tend to be trained to predict as many salient instances as possible. However, they may predict multiple instances from a salient instance of self-contrast (*e.g.*, rows 1, 2, 5 and 6) and sometimes miss salient instances of solid colors (*e.g.*, rows 4 and 6). In comparison, our method can successfully detect salient instances and delineate their boundaries, with the help of learning intra-instance saliency coherence and cross-instance saliency contrast.

4.4. Ablation Study

Ablation Study of SCoCo. We first study the effectiveness of the proposed GBA and LAS modules, and the loss functions $\mathcal{L}_{\mathcal{S}}$ and $\mathcal{L}_{\mathcal{I}}$ used in them. The mAP$_{50}$ results in Table 2 show that solely applying either scene-level (GBA + $\mathcal{L}_{\mathcal{S}}$) or instance-level (LAS + $\mathcal{L}_{\mathcal{I}}$) saliency contrast learning may produce a performance decrease of 2-3%, and the performance drops by around 4% if we use the GBA or LAS module only. Combining all of them produces the best performance, which verifies the necessities of our designs.

Table 2. Ablation study of the GBA and LAS modules, and the corresponding losses, $\mathcal{L}_{\mathcal{S}}$ and $\mathcal{L}_{\mathcal{I}}$, on the SIS [46] test set using mAP metrics. Best performances are marked in **bold**.

Method	GBA	$\mathcal{L}_{\mathcal{S}}$	LAS	$\mathcal{L}_{\mathcal{I}}$	mAP$_{50}$	mAP$_{70}$
SCoCo					64.7%	52.6%
	✓				67.0%	53.5%
	✓	✓			68.4%	54.1%
			✓		67.4%	53.1%
			✓	✓	69.0%	54.4%
	✓	✓	✓	✓	**71.4%**	**55.7%**

Visualization of the GBA Module. Figure 7 visualizes the scene-level salinecy map learned by the GBA module and the corresponding SID result. We can see that without the GBA module, the model tends to be distracted by high-contrast background regions, such as the shadow, and identifies them as salient regions. In contrast, using the GBA module can successfully suppress such noise by learning proper saliency thresholds.

Visualization of the LAS Module. In Figure 8, we investigate the influence of our LAS module by visualizing feature correlations to points (purplish dots) in individual instances. We can see that the two pigs have non-uniform white and

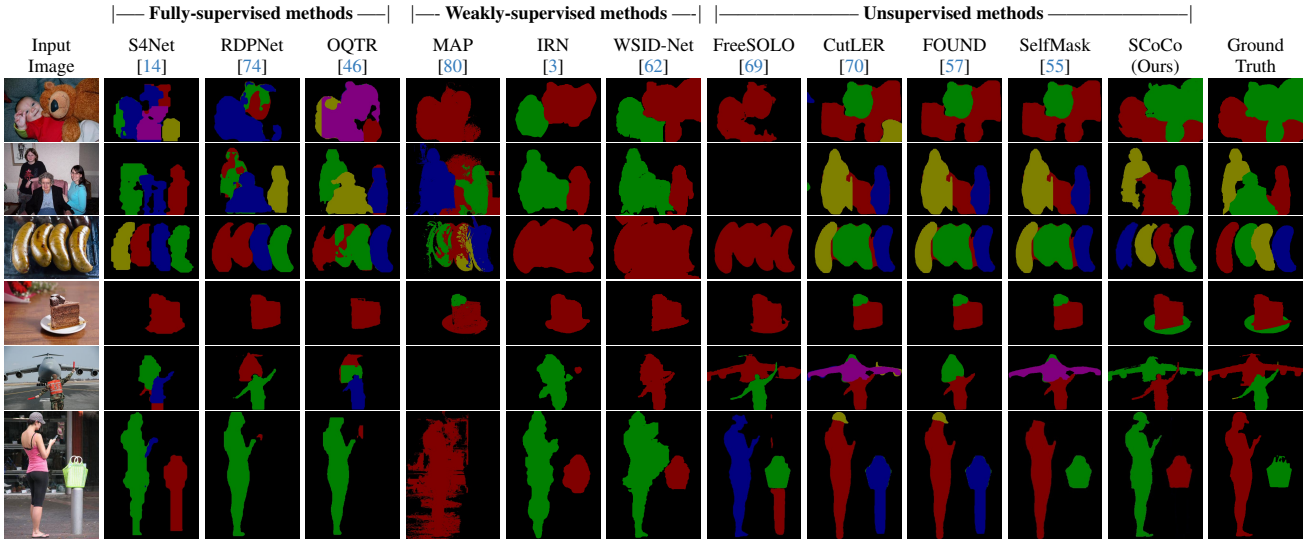


Figure 6. Qualitative comparison with existing state-of-the-art methods and baselines. Each color represents a unique salient instance.

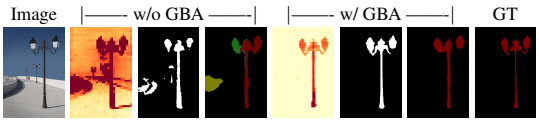
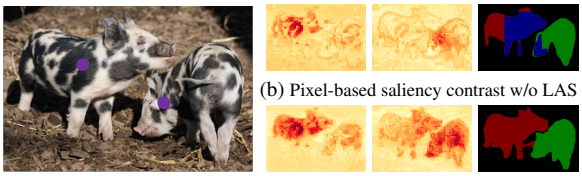


Figure 7. Visualization of the GBA module.



(a) Input Image (b) Pixel-based saliency contrast w/o LAS (c) Pixel-based saliency contrast w/ LAS

Figure 8. Visualization of the LAS module.

black colors on their bodies, causing our model to learn incomplete saliency coherence for each pig and produce over-detection results without the LAS module (b). With the LAS module, our model can separate the two pigs successfully by learning saliency contrast between instances while maintaining intra-instance saliency coherence.

Backbone Initialization Choices. Finally, we study the alternatives of self-supervised methods for initializing our ViT backbone. As shown in Table 3, in comparison to MoCo-v3 [18] and SWaV [7], we choose DINO [8] to initialize our backbone as it can provide better performances from its learned feature similarities. In addition, we select ViT-S as our backbone since ViT-B can only introduce minor improvements with larger parameters.

5. Conclusion

In this paper, we have studied a new task, *unsupervised salient instance detection*. We have observed that self-supervised transformer features exhibit intra-region saliency coherence and different levels of inter-region saliency contrast, which can help differentiate salient instances. With this observation, we have proposed SCoCo, with two novel designs, to model such regional saliency coherence and contrast information. First, we propose a global background adaptation (GBA) module and a scene-level contrastive loss to find saliency thresholds for salient objects w.r.t. diverse background contents. Second, we propose a locality-aware similarity (LAS) module and an instance-level contrastive loss to model in-region saliency coherence and cross-region saliency contrasts for detecting salient instances. Extensive experiments have verified the effectiveness of our method against SOTA methods.

Table 3. Quantitaive analysis of backbone initialization choices. Best performances are marked in **bold**.

Method	Init.	Arch.	mAP_{50}	mAP_{70}
SCoCo	MoCo-v3 [18]	ViT-S	70.5%	54.9%
	SWaV [7]	ViT-S	70.2%	54.3%
	DINO [8]	ViT-S	71.4%	**55.7%**
	DINO [8]	ViT-B	**71.9%**	55.6%

Our method does have limitations. We note that SCoCo may not be able to detect small salient instances as it learns from high-level, low-resolution features. As shown in Figure 9, the person standing on the speedboat is too small to be detected by SCoCo. As a future work, it would be interesting to study multi-granularity salient feature similarity and contrast modeling to detect multi-scale salient instances.

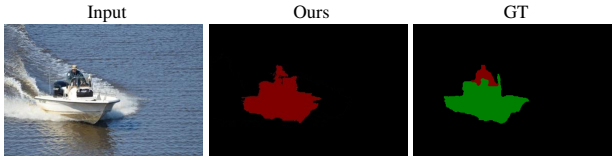


Figure 9. SCoCo may fail to detect small salient instances.

References

[1] Kfir Aberman, Junfeng He, Yossi Gandelsman, Inbar Mosseri, David E Jacobs, Kai Kohlhoff, Yael Pritch, and Michael Rubinstein. Deep saliency prior for reducing visual distraction. In *CVPR*, 2022. 1

[2] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009. 3

[3] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, 2019. 6, 7, 8, 1

[4] Codruta Orniana Ancuti, Cosmin Ancuti, and Phillipe Bekaert. Enhancing by saliency-guided decolorization. In *CVPR*, 2011. 1

[5] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *CVM*, 2019. 3

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 5

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 2020. 8

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 6, 8

[9] Zixuan Chen, Huajun Zhou, Jianhuang Lai, Lingxiao Yang, and Xiaohua Xie. Contour-aware loss: Boundary-aware learning for salient object segmentation. *IEEE TIP*, 2020. 3

[10] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 2014. 3

[11] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *CVPR*, 2015. 3

[12] Hisham Cholakkal, Guolei Sun, Fahad Shahbaz Khan, and Ling Shao. Object counting and instance segmentation with image-level supervision. In *CVPR*, 2019. 6, 7

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 6

[14] Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. S4net: Single stage salient-instance segmentation. In *CVPR*, 2019. 2, 6, 7, 8, 1

[15] Shuyong Gao, Wei Zhang, Yan Wang, Qianyu Guo, Chenglong Zhang, Yangji He, and Wenqiang Zhang. Weakly-supervised salient object detection using point supervision. In *AAAI*, 2022. 3

[16] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 6

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 5, 6

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 8

[19] Brian Hu, Bhavan Vasu, and Anthony Hoogs. X-mir: Explainable medical image retrieval. In *WACV*, 2022. 1

[20] Lai Jiang, Yifei Li, Shengxi Li, Mai Xu, Se Lei, Yichen Guo, and Bo Huang. Does text attract attention on e-commerce images: A novel saliency prediction dataset and method. In *CVPR*, 2022. 1

[21] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1, 6

[22] Yun Yi Ke and Takahiro Tsubono. Recursive contour-saliency blending network for accurate salient object detection. In *WACV*, 2022. 3

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012. 6

[24] Duo Li, Jie Hu, Changhu Wang, Xiangtai Li, Qi She, Lei Zhu, Tong Zhang, and Qifeng Chen. Involution: Inverting the inherence of convolution for visual recognition. In *CVPR*, 2021. 4

[25] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *CVPR*, 2017. 1, 2, 6, 7

[26] Guanbin Li, Yuan Xie, and Liang Lin. Weakly supervised salient object detection using image labels. In *AAAI*, 2018. 3

[27] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, 2013. 3

[28] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022. 6, 7

[29] Ming Liang and Xiaolin Hu. Recurrent convolutional neural network for object recognition. In *CVPR*, 2015. 3

[30] Zijian Liang, Pengjie Wang, Ke Xu, Pingping Zhang, and Rynson WH Lau. Weakly-supervised salient object detection on light fields. *IEEE TIP*, 2022. 3

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 6

[32] Xiangru Lin, Ziyi Wu, Guanqi Chen, Guanbin Li, and Yizhou Yu. A causal debiasing framework for unsupervised salient object detection. In *AAAI*, 2022. 1, 3

[33] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, 2019. 3

[34] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016. 3

[35] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018. 3

[36] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *ICCV*, 2021. 3

[37] Nian Liu, Wangbo Zhao, Ling Shao, and Junwei Han. Scg: Saliency and contour guided salient instance segmentation. *IEEE TIP*, 2021. 2, 5, 6, 7

[38] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE TPAMI*, 2010. 3

[39] Yudong Liu, Yongtao Wang, Siwei Wang, TingTing Liang, Qijie Zhao, Zhi Tang, and Haibin Ling. Cbnet: A novel composite backbone network architecture for object detection. In *AAAI*, 2020. 3

[40] Yuxuan Liu, Pengjie Wang, Ying Cao, Zijian Liang, and Rynson WH Lau. Weakly-supervised salient object detection with saliency bounding boxes. *IEEE TIP*, 2021. 3

[41] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *CVPR*, 2017. 6, 7

[42] Ningning Ma, Xiangyu Zhang, Ming Liu, and Jian Sun. Activate or not: Learning customized activation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8032–8042, 2021. 4

[43] S Mahdi H Miangoleh, Zoya Bylinskii, Eric Kee, Eli Shechtman, and Yağiz Aksoy. Realistic saliency guided image enhancement. In *CVPR*, 2023. 1

[44] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction via self-supervision. *NeurIPS*, 2019. 1, 3

[45] Jialun Pei, He Tang, Chao Liu, and Chuanbo Chen. Salient instance segmentation via subitizing and clustering. *Neurocomputing*, 2020. 2, 6, 7

[46] Jialun Pei, Tianyang Cheng, He Tang, and Chuanbo Chen. Transformer-based efficient salient instance segmentation networks with orientative query. *IEEE TMM*, 2022. 2, 5, 6, 7, 8, 1

[47] Jialun Pei, Tao Jiang, He Tang, Nian Liu, Yueming Jin, Deng-Ping Fan, and Pheng-Ann Heng. Calibnet: Dual-branch cross-modal calibration for rgb-d salient instance segmentation. *arXiv preprint arXiv:2307.08098*, 2023. 2

[48] Guohao Peng, Jun Zhang, Heshan Li, and Danwei Wang. Attentional pyramid pooling of salient visual residuals for place recognition. In *ICCV*, 2021. 1

[49] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012. 3

[50] Yongri Piao, Jian Wang, Miao Zhang, and Huchuan Lu. Mfnet: Multi-filter directive network for weakly supervised salient object detection. In *ICCV*, 2021. 3

[51] Pedro OO Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *NeurIPS*, 2015. 7

[52] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019. 3

[53] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017. 4

[54] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 2000. 5

[55] Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *CVPR Workshop*, 2022. 1, 3, 4, 6, 7, 8

[56] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *BMVC*, 2021. 3, 5

[57] Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonín Vobeckỳ, Éloi Zablocki, and Patrick Pérez. Unsupervised object localization: Observing the background to discover objects. In *CVPR*, 2023. 1, 3, 4, 6, 7, 8

[58] Avishek Siris, Jianbo Jiao, Gary KL Tam, Xianghua Xie, and Rynson WH Lau. Scene context-aware salient object detection. In *ICCV*, 2021. 3

[59] Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *CVPR*, 2019. 4

[60] Chang Tang, Xinzhong Zhu, Xinwang Liu, and Pichao Wang. Salient object detection via recurrently aggregating spatial attention weighted cross-level deep features. In *ICME*, 2019. 3

[61] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Weakly-supervised salient instance detection. In *BMVC*, 2020. 2

[62] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to detect instance-level salient objects using complementary image labels. *IJCV*, 2022. 1, 2, 6, 7, 8

[63] Huy V Vo, Francis Bach, Minsu Cho, Kai Han, Yann LeCun, Patrick Pérez, and Jean Ponce. Unsupervised image matching and object discovery as optimization. In *CVPR*, 2019. 3

[64] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *ECCV*, 2020.

[65] Van Huy Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. *NeurIPS*, 2021. 3

[66] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, 2016. 3

[67] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 3

[68] Xiang Wang, Huimin Ma, Xiaozhi Chen, and Shaodi You. Edge preserving and multi-scale contextual neural network for salient object detection. *TIP*, 2017. 3

[69] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. Freesolo: Learning to segment objects without annotations. In *CVPR*, 2022. 3, 5, 6, 7, 8, 1

[70] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *CVPR*, 2023. 3, 5, 6, 7, 8, 1

[71] Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L Crowley, and Dominique Vaufreydaz. Tokencut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. *arXiv:2209.00383*, 2022. 3, 5

[72] Yifan Wang, Wenbo Zhang, Lijun Wang, Ting Liu, and Huchuan Lu. Multi-source uncertainty mining for deep unsupervised saliency detection. In *CVPR*, 2022. 1, 3

[73] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *CVPR*, 2020. 3

[74] Yu-Huan Wu, Yun Liu, Le Zhang, Wang Gao, and Ming-Ming Cheng. Regularized densely-connected pyramid network for salient instance segmentation. *IEEE TIP*, 2021. 1, 6, 7, 8

[75] Pengxiang Yan, Ziyi Wu, Mengmeng Liu, Kun Zeng, Liang Lin, and Guanbin Li. Unsupervised domain adaptive salient object detection through uncertainty-aware pseudo-label learning. In *AAAI*, 2022. 3

[76] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 3

[77] Siyue Yu, Bingfeng Zhang, Jimin Xiao, and Eng Gee Lim. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *AAAI*, 2021. 3

[78] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, and Yizhou Yu. Multi-source weak supervision for saliency detection. In *CVPR*, 2019. 3

[79] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *ICCV*, 2017. 1, 3

[80] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Unconstrained salient object detection via proposal subset optimization. In *CVPR*, 2016. 2, 6, 7, 8, 1

[81] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *CVPR*, 2018. 1, 3

[82] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *CVPR*, 2020. 3

[83] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *CVPR*, 2018. 3

[84] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, 2018. 3

[85] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, 2019. 3

[86] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *CVPR*, 2020. 3

[87] Huajun Zhou, Peijia Chen, Lingxiao Yang, Xiaohua Xie, and Jianhuang Lai. Activation to saliency: Forming high-quality labels for unsupervised salient object detection. *IEEE TCSVT*, 2022. 1, 3

[88] Huajun Zhou, Bo Qiao, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Texture-guided saliency distilling for unsupervised salient object detection. In *CVPR*, 2023. 1, 3

[89] Huajun Zhou, Bo Qiao, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Texture-guided saliency distilling for unsupervised salient object detection. In *CVPR*, 2023. 3

[90] Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, and Ming-Hsuan Yang. Decoupled dynamic filter networks. In *CVPR*, 2021. 4

[91] Zikun Zhou, Wenjie Pei, Xin Li, Hongpeng Wang, Feng Zheng, and Zhenyu He. Saliency-associated object tracking. In *ICCV*, 2021. 1

[92] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *CVPR*, 2014. 3

Unsupervised Salient Instance Detection

Supplementary Material

Paper ID: 8768

In this supplementary material, we provide more quality comparisons between the proposed SCoCo and state-of-the-art methods and baselines to further support the visual comparison in our main paper.

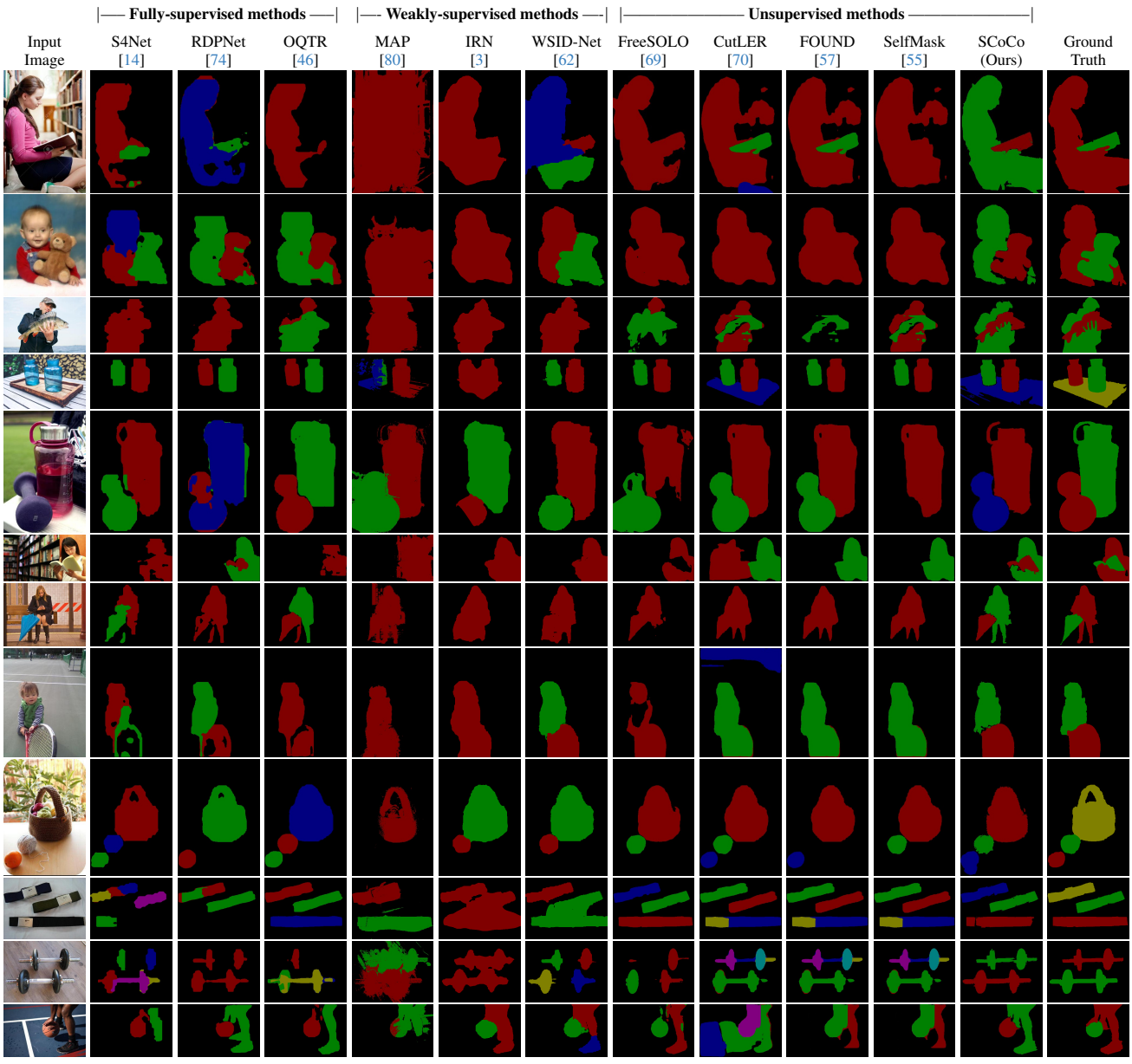


Figure S1. Qualitative comparison with existing state-of-the-art methods and baselines. Each color represents a unique salient instance.