Article

CoSOV1Net: Cone- and Spatial-Opponency Primary Visual Cortex-Inspired Neural Network for Lightweight Salient Object Detection

Didier Ndayikengurukiye [1,†,*] and Max Mignotte [1,‡]

[1] Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Québec, Canada;
[†] didier.ndayikengurukiye@umontreal.ca
[‡] mignotte@iro.umontreal.ca
[*] Correspondence: didier.ndayikengurukiye@umontreal.ca

Abstract: Computer vision models of salient object detection attempt to mimic the ability of the human visual system to select relevant objects in images. To this end, the development of deep neural networks on high-end computers has recently made it possible to achieve high performance. However, it remains a challenge to develop deep neural network models of the same performance for devices with much more limited resources. In this work, we propose a new approach for a lightweight salient object detection neural network model, inspired by the cone and spatial opponent processes of the primary visual cortex (V1), that inextricably link color and shape in human color perception. Our proposed model, namely CoSOV1net, is trained from scratch, without using backbones from image classification or other tasks. Experiments, on the most widely used and challenging datasets for salient object detection, show that CoSOV1Net achieves competitive performance (i.e. $F_\beta = 0.931$ on the ECSSD dataset) with state-of-the-art salient object detection models, while having low number of parameters ($1.14M$), low FLOPS ($1.4G$) and high FPS (211.2) on GPU (nvidia Geforce RTX 3090 TI) compared to the state-of-the-art in the salient object detection or lightweight salient object detection task. Thus, CoSOV1net turns out to be a lightweight salient object detection that can be adapted to mobile environments and resource-constrained devices.

Keywords: lightweight salient object detection; salient object detection; object detection; lightweight neural network; color opponent; cone-opponent; double-opponent; vision sensing

1. Introduction

The human visual system (HVS) has the ability to select, among the large amount of information received, which is relevant and to process in detail only the relevant one. This relevant information in an image is called salient objects [1]. The salient object detection models in computer vision try to mimic this phenomenon by detecting and segmenting salient objects in images.

The salient object detection is an important task given its many applications in computer vision such as object tracking, recognition and detection [2] , advertisements optimization [3], image/video compression [4], image correction [5], analysis of iconographic illustrations [6], images retrieval [7], aesthetic evaluation [8], image quality evaluation [9], image retargeting [10], image editing [11], image collage [12], to name a few. Thus, it has been the subject of intensive research in recent years and is still being investigated [13].

Salient object detection models generally fall into two categories, namely conventional and deep learning-based models, which differ by the feature extraction process. The first use hand-crafted features, while the latter use features learned from a neural network. Thanks to the powerful representation learning methods, deep learning-based salient object detection models have recently shown superior performance over conventional models [13,14].

The high performance of deep learning-based salient object detection models is undeniable, however, they are also generally heavy if we consider their number of parameters and memory occupied in addition to their high computational cost and slow detection

check for updates

Citation: Ndayikengurukiye, D.; Mignotte, M. CoSOV1Net: Cone- and Spatial-Opponency Primary Visual Cortex-Inspired Neural Network for Lightweight Salient Object Detection. *Preprints* **2023**, *1*, 0. https://doi.org/

 doi:10.20944/preprints202306.0899.v1

speed. This makes these models less practical for resource-limited vision sensors or mobile devices that have many constraints on their memory, computational capabilities and for real-time applications [15,16]. Hence, the need for lightweight salient object detection models, whose performance is comparable to state-of-the-art models, with the advantage of being deployed on resource-limited vision sensors or mobile devices and having a detection speed allowing them to be used in real time applications.

The authors of these existing models of lightweight salient object detection authors have used different methodologies such as imitating primate hierarchical visual perception [17], human attention mechanism [16,17], etc.

In this work, we propose an original approach for a new lightweight neural network model of salient object detection that can be therefore adapted to mobile environments and limited or resources-constrained devices with the following additional interesting properties of being able to be trained from scratch, without having to use backbones developed from image classification tasks and with few parameters but while having comparable performance with state-of-the-art models.

Given that detecting salient objects is a capability of the human visual system, and that a normal human visual system does this quickly and correctly, we used images or scenes encoding mechanism research advances in neuroscience especially for the human visual system early stage [18–20]. Our strategy in this model is therefore inspired by two neuroscience discoveries in human color perception, namely:

1. the color-opponency encoding in the HVS (Human Visual System) early stage [21–24];
2. the fact that the color and pattern are linked inextricably in human color perception [18,25].

Inspired by these neuroscience discoveries, we propose a Cone- and Spatial-Opponency Primary Visual cortex (CoSOV1) module that extracts the features at the spatial level and between the color channels at the same time to integrate color in the patterns. This process is applied first on opposing color pairs channels two by two and then to grouped feature maps through our deep neural network. Thus, based on the CoSOV1 module, we build a novel lightweight encoder-decoder deep neural network for salient object detection: CoSOV1Net. Our proposed neural network CoSOV1Net has only $1.14M$ parameters while having comparable performance with state-of-the-art salient object detection models. CoSOV1Net predicts salient maps on a speed of $4.4fps$ on an Intel CPU, i7-11700F, and $211.2fps$ on a NVIDIA Geforce RTX 3090 TI GPU for 384× 384 images; and has a low $FLOPS = 1.4G$. Therefore, CoSOV1net is a lightweight salient object detection that can be adapted for mobile environments and limited resources devices.

Our contribution is threefold:

- we propose a novel approach to extract features from opposing color pairs in a neural network to exploit the strength of color opponent principle from human color perception. This approach permits to accelerate neural network learning;
- we propose a strategy to integrate color in patterns in a neural network by extracting features locally and between color channels at the same time in successively grouped feature maps that results in reducing the number of parameters and the depth of that neural network, while keeping good performance;
- we propose a lightweight salient object detection neural network architecture based on the proposed approach for learning opposing color pairs along with the strategy of integrating color in patterns. This lightweight salient object detection neural network has few parameters while having performance comparable to the state-of-the-art methods

The rest of this work is organized as follows: Section 2 presents some lightweight models related to this approach ; Section 3 presents our proposed lightweight salient object detection model; Section 4 describes the datasets used, evaluation metrics, our experimental results and the comparison of our model with state-of-the-art models; Section 5 discusses our results ; Section 6 concludes this work.

 doi:10.20944/preprints202306.0899.v1

2. Related work

Many salient object detection models have been proposed and most of the influential advances in image-based salient object detection have been reviewed by Gupta *et al.* [13]. Herein, we present some conventional models and lightweight neural network models related to this approach.

2.1. Lightweight Salient Object Detection

In recent years, lightweight salient object detection models have been proposed with different strategies and architectures. Qin *et al.* [26] designed U^2net a lightweight salient object detection with a two-level nested Unet[27] neural network able to capture more contextual information from different scales thanks to the mixture of receptive fields of different sizes. Other models are based on streamlined architecture to build lightweight deep neural networks. MobileNets [28,29] and ShuffleNets [30,31] with their variants are among the latter models. MobileNets [28] uses architecture based on depth-wise separable convolutions. ShuffleNets [30] used architecture based on point-wise group convolution and channel shuffle and depth-wise convolution to greatly reduce computation cost while maintaining accuracy. Other authors have been inspired by primate or human visual system process. Thus, Liu *et al.* [17] designed the HVPNet, a lightweight salient object detection network, based on a hierarchical visual perception (HVP) module which mimic the primate visual cortex for hierarchical perception learning. Liu *et al.* [16] were inspired by human perception attention mechanism in designing SAMNet a lightweight salient object based on a stereoscopically attentive multiscale (SAM) module that adopts a stereoscopic attention mechanism for effective and efficient multi-scale learning.

2.2. Opponent Color models

The color opponency, which is a human color perception propriety, has inspired many authors who have defined channels or feature maps to tackle their image processing tasks. Frintrop *et al.* [32] used three opponent channels RG, BY and I to extract features for their salient object detection model.

To extract features for salient object detection, Ndayikengurukiye *et* Mignotte [1], used nine (9) opponent channels for RGB color space (RR: Red-Red, RG: Red-Green, RB: Red-Blue, GR: Green-Red, GG: Green-Green, GB: Green-Blue, BR: Blue-Red, BG: Blue-Green and BB: Blue-Blue) with a non linear combination thanks to the OCLTP (opponent color local ternary pattern) texture descriptor which is an extension of the OCLBP (opponent color local binary pattern)[33,34] and Fastmap [35] which is a fast version of MDS (Multi-dimensional Scaling).

Most authors apply the opponent color mechanism to the input image color space channels and not on the resulting feature maps. However, Jain *et* Healey [36] used opponent features computed from Gabor filter outputs. These authors compute opponent features by combining information across different spectral bands at different scales obtained by Gabor filters for color texture recognition [36]. Yang *et al.* [37] proposed a framework based on the color-opponent mechanisms of color-sensitive double-opponent (DO) cells in the human visual system primary visual cortex (V1) in order to combine brightness and color to maximize the boundary detection reliability in natural scenes.

In this work, we propose a model inspired by the human visual system but different from other models because our model uses the primary visual cortex (V1) cone- and spatial-opponency principle to extract features at channels spatial level and between color channels at the same time to integrate color into patterns in a manner allowing lightweight deep neural network design.

3. Materials and Methods

3.1. Introduction

Our model for tackling the lightweight salient object detection challenge is inspired by the human visual system (HVS) early visual color process especially its cone-opponency

 doi:10.20944/preprints202306.0899.v1

and spatial-opponency in the primary visual cortex (V1). Indeed, the human eye retina (the retina is located in the inner surface of the human eye) has two types of photoreceptors namely rods and cones. Rods are responsible for monochromatic vision under low levels of illumination while cones are responsible for color vision at normal levels of illumination. There are three classes of cones: the L, M and S. When light is absorbed by cone photoreceptors, the L cones absorbs long wavelength visible light, the M cones the middle ones and the S cones the short wavelength [22,23,25].

The cone signals are then processed by single-opponent retina ganglion cells. The single-opponent operates an antagonistic comparison of the cone signals [21,23,24,38]:

- L - M opponent for Red - Green;
- S - (L + M) opponent for Blue - Yellow.

The Red - Green as well as the Blue - Yellow signals are carried by specific cells (cells for Red - Green and different cells for Blue - Yellow) through the lateral geniculate nucleus (LGN) to primary visual cortex (V1).

Shapley [25] and Shapley *et* Hawken [18] showed that the primary visual cortex (V1) plays an important role in color perception through the combined activity of two kinds of color-sensitive cortical neurons namely single-opponent and double-opponent cells. Single-opponent cells in V1 operate in the same manner as those of retina ganglion cells and could provide neuronal signals that could be used for estimating the color of the illumination [25]. Double-opponent cells in V1 compare cone signals across space as well as between cones [19,20,22,25]. Double-opponent thus have two oppenencies: spatial-opponency and cone-opponency. These properties permit them to be sensitive to color edges and color spatial patterns. They are thus able to link color and pattern inextricably in human color perception [18,25].

As the primary visual cortex (V1) is known to play a major role in visual color perception as highlighted above, we propose in this work a deep neural network based on the primary visual cortex (V1) to tackle lightweight salient object detection challenge. We especially use two neuroscience discoveries in human color perception, namely:

1. the color-opponency encoding in the HVS early stage;
2. the fact that the color and pattern are linked inextricably in human color perception.

These two discoveries in neuroscience inspired us to design a neural network architecture for lightweight salient object detection, which hinges on two main ideas. First, at the beginning of the neural network, our model opposes color channels two by two by grouping them (R-R, R-G, R-B, G-G, G-B, B-B) and extract from each channels pair the features at the channels spatial level and between the color channels at the same time to integrate color in patterns. So, instead of doing a subtractive comparison or an OCLTP (opponent color linear ternary pattern) like Ndayikengurukiye *et* Mignotte [1], we let the neural network learn the features that represent the comparison of the two color pairs. Second, this idea of grouping and then extracting the features at the channels spatial level and between the color channels at the same time is applied on feature maps at each neural network level until the saliency maps are obtained. This process allows the proposed model to mimic the human visual system capability of linking inextricably color and pattern in color perception [18,25].

It is this idea that differentiates our model from other models that use depth-wise convolution first and after point-wise convolution [28,29] to extract features at each individual color channel level (or feature map) first not through a group of color channels (or feature maps) at same time as our model does. This idea differentiates also our model from models that combine a group of color channels (or feature maps) pixel by pixel first and apply depth-wise convolution after[30,31]. The idea of grouping color channels in pairs (or feature maps groups) differentiates our model from models that, while extracting features

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

at color channels spatial level and between color channels at the same time, consider all color channels (or feature maps) as a single group.

Our model takes into account non-linearities in the image at the beginning as well as through our neural network. For this purpose, we use an encoder-decoder neural network type whose core is a module that we called CoSOV1 (Cone- and Spatial-Opponency Primary Visual Cortex) module.

3.2. CoSOV1 : Cone- and Spatial-Opponency Primary Visual Cortex module

The CoSOV1 (Cone- and Spatial-Opponency Primary Visual Cortex) module is composed of two parts (see Figure 1).

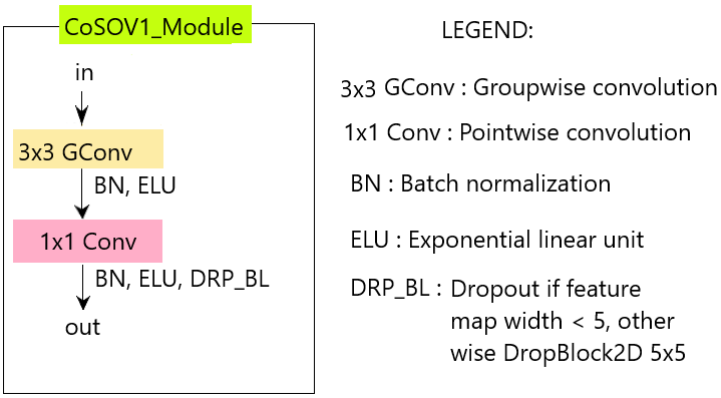

Figure 1. The CoSOV1 (Cone- and Spatial-Opponency Primary Visual Cortex) module is the core of our neural network model.

In the first part, input color channels (or input feature maps) are split into groups of equal depth. Convolution (3×3) operations are then applied on each group of channels (or feature maps) in order to extract features from each group as opposing color channel (or opposing feature maps). This is done thanks to a set of filters that convolve the group of color channels (or feature maps). Each filter is applied to the color channels (or input feature maps) through a convolution operation which detects local features at all locations on the input. Let $\mathcal{I}^g \in \mathbb{R}^{\mathcal{W}\times\mathcal{H}\times S}$ be a input group of feature maps where $\mathcal{W}$ and $\mathcal{H}$ are respectively the width and the height of each group's feature map and $W \in \mathbb{R}^{3\times3\times S}$, a filter with learned weights, S being the depth of each group or the number of the channels in each group g. The output feature map $\mathcal{O}^g \in \mathbb{R}^{\mathcal{W}\times\mathcal{H}}$ for this group $g \in \{1,\ldots,\mathcal{G}\}$ has pixel value in (k,l) position defined as follow:

$$\mathcal{O}^g_{k,l} = \sum_{s=1}^{S}\sum_{i=0}^{2}\sum_{j=0}^{2} W_{i,j,s}\mathcal{I}^g_{k+i-1,l+j-1,s} \tag{1}$$

The weights matrix $W \in \mathbb{R}^{3\times3\times S}$ is the same across the whole group of channels or feature maps. So, each resulting output feature map represents a particular feature at all locations in the input color channels (or input feature maps) [39]. We call the 3×3 convolution on grouped channels (or grouped feature maps) a group-wise convolution. The zero padding is applied during the convolution process to keep the input channels size for the output feature maps. After the group-wise convolution, we apply the batch normalization transform which is known to enable faster and more stable training of deep neural networks [40,41]. Let $\mathfrak{B} = \{X_1,\ldots,X_m\}$ be a mini-batch that contains m examples from the dataset $\{X_1,\ldots,X_m\}$, the mini-batch mean is

$$\mu_{\mathfrak{B}} = \frac{1}{m}\sum_{i=1}^{m} X_i$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

and the mini-batch variance is

$$\sigma_{\mathfrak{B}}^2 = \frac{1}{m}\sum_{i=1}^{m}(X_i - \mu_{\mathfrak{B}})^2$$

The batch normalization transform $BN_{\gamma,\beta} : \{X_1, \ldots, X_m\} \longrightarrow \{Y_1, \ldots, Y_m\}$ (γ and β being parameters to be learned):

$$Y_i = \gamma\widehat{X_i} + \beta \tag{2}$$

where $i \in \{1, \ldots, m\}$ and

$$\widehat{X_i} = \frac{X_i - \mu_{\mathfrak{B}}}{\sqrt{\sigma_{\mathfrak{B}}^2 + \epsilon}}$$

and ϵ is a very small constant to avoid division by zero.

In order to take into account the non-linearities present in the color channels input (or feature maps input), given that group-wise is a linear transformation, the batch normalization is followed by a non-linear function, Exponential Linear Unit (ELU) defined as follows:

$$ELU(x) = \begin{cases} x & \text{if } x \geq 0, \\ \alpha \times (\exp(x) - 1.) & \text{otherwise} \end{cases}$$

$\alpha = 1$ by default.

The non-linear function which is the activation function is placed after batch normalization as recommended by Chollet [42].

The second part of the module searches the best representation of the obtained feature maps. It is similar to the first part of the module except for group-wise convolution which is replaced by the point-wise convolution but the input feature maps for the point-wise convolution in this model are not grouped. The point-wise convolution allows us to learn the filters weights and thus obtain feature maps that best represent the input channels (or input feature maps) for the salient object detection task while having few parameters.

Let $\mathcal{O} \in \mathbb{R}^{\mathcal{W}\times\mathcal{H}\times M}$ be the output of the first part of the module, M being the number of feature maps in this output, $\mathcal{W}$, $\mathcal{H}$ the width and the height respectively. Let a filter of the learned weights $V \in \mathbb{R}^M$ and $\mathcal{FM} \in \mathbb{R}^{\mathcal{W}\times\mathcal{H}}$, its output feature map by a point-wise convolution. Its pixel's value $\mathcal{FM}_{k,l}$ in (k, l) position is:

$$\mathcal{FM}_{k,l} = \sum_{m=1}^{M} V_m \mathcal{O}_{k,l,m} \tag{3}$$

Thus, $V \in \mathbb{R}^M$ is a vector of learned weights which, to the input feature maps $\mathcal{O} \in \mathbb{R}^{\mathcal{W}\times\mathcal{H}\times M}$ associates a feature map $\mathcal{FM} \in \mathbb{R}^{\mathcal{W}\times\mathcal{H}}$ which is the best representation of the input feature maps $\mathcal{O} \in \mathbb{R}^{\mathcal{W}\times\mathcal{H}\times M}$. The point-wise convolution in this module uses many filters and thus it outputs many feature maps that are the best representation of the input feature maps $\mathcal{O}$. As point-wise convolution is a linear combination, we apply again a batch normalization followed by a exponential linear unit function (ELU) on the feature map $\mathcal{FM}$, to get the best representation of the input feature maps for the learned weights $V \in \mathbb{R}^M$ that takes into account non-linearities in the feature maps $\mathcal{O} \in \mathbb{R}^{\mathcal{W}\times\mathcal{H}\times M}$.

Our scheme is different from depth-wise separable convolution in that depth-wise convolution does not use the convolution of a group of channels but each channel individually [28,43]

In addition, after the non-linear function, noise is injected in the resulting feature maps during the neural network learning stage thanks to dropout process (but not in the prediction stage) to facilitate the learning process. We used in this model the dropblock [44] if the width of the feature map is greater than 5 and the common dropout [45] otherwise.

The CoSOV1 module allows to have few parameters but also good performance for our neural network.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

3.3. CoSOV1Net neural network model architecture

Our proposed model is built on the CoSOV1 module (see Figure 1). It is a neural network of the U-net encoder-decoder type [27] and is illustrated in Figure 2. Thus, our model consists of three main blocks:

1. The input RGB color channel pairing;
2. The encoder;
3. The decoder.

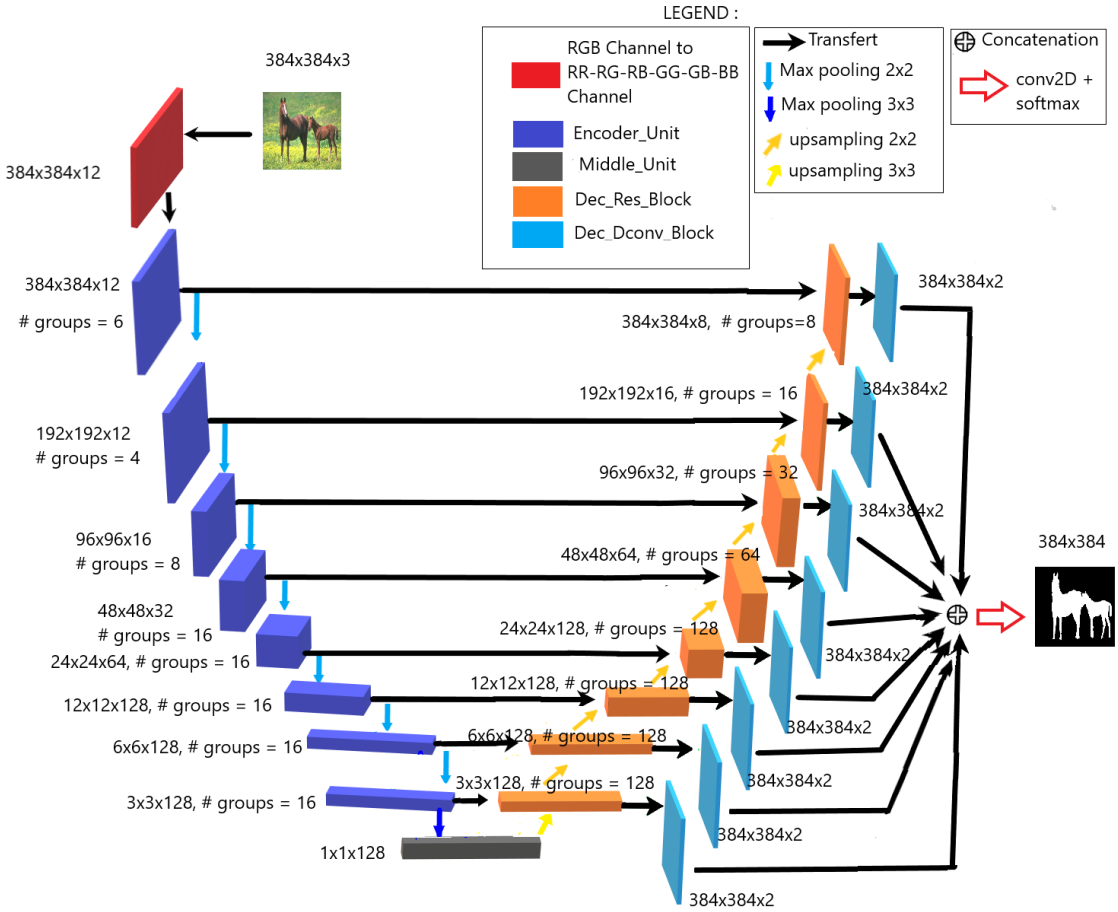

Figure 2. Our model CoSOV1 neural network architecture: (a) Red block (input RGB color channel pairing for opposing color pairs; (b) Violet blocks for the encoder; (c) Grey block for the middle of the model; (d) Pink blocks for the decoder residual block; (e) Blue blocks for the deconvolution and upsampling of all scales feature maps to the initial scale.

3.3.1. Input RGB color channel pairing

At this stage, the input RGB image is paired in 6 opposing color channels pairs R-R, R-G, R-B, G-G, G-B, B-B [1,33,46]. These pairs are then concatenated which give 12 channels R,R,R,G,R,B,G ,G,G,B,B,B as illustrated in Figure 3. This is the step for choosing the color channels to oppose. The set of color channels concatenated are then fed to the encoder.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

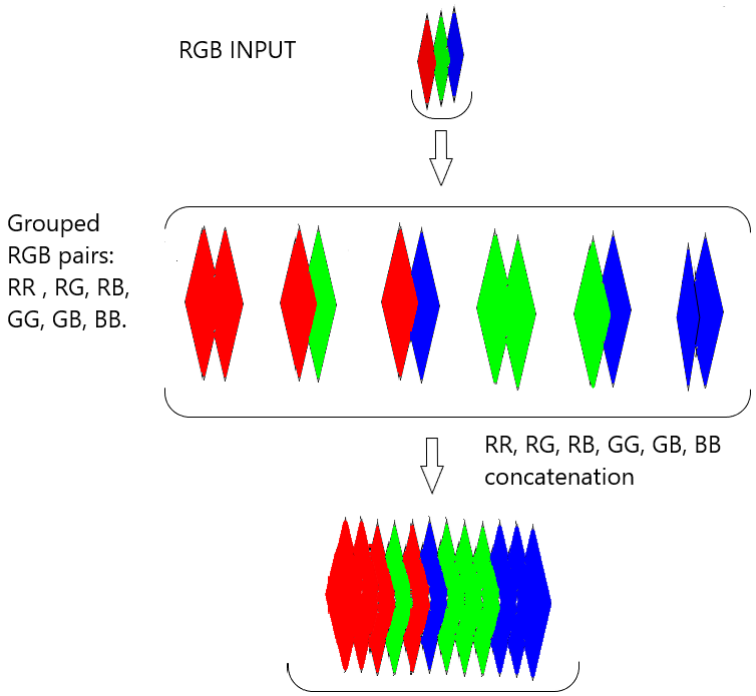

Figure 3. Input RGB color image transformed in 6 opposing color channels pairs, these are then concatenated to get 12 color channels.

3.3.2. Encoder

The Encoder, in our proposed neural network model, is a convolutional neural network (CNN) [47] where an Encoder Unit (see Figure 2) is repeated eight times. Each Encoder Unit is followed by a max pooling (2×2) with strides=2 except for the 8^{th} neural network level where the max pooling is 3×3 with strides=3 (the max pooling is a downsampling operation, like a filtering with a maximum filter). While the size of each feature map is reduced by half, the depth of the feature maps is doubled except for the first level where it is kept at 12 and the last two levels where it is kept at 128 to have few parameters.

The Encoder Unit (see Figure 4 (a)) is composed of a residual block (Figure 4 (b)) repeated three (3) times.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

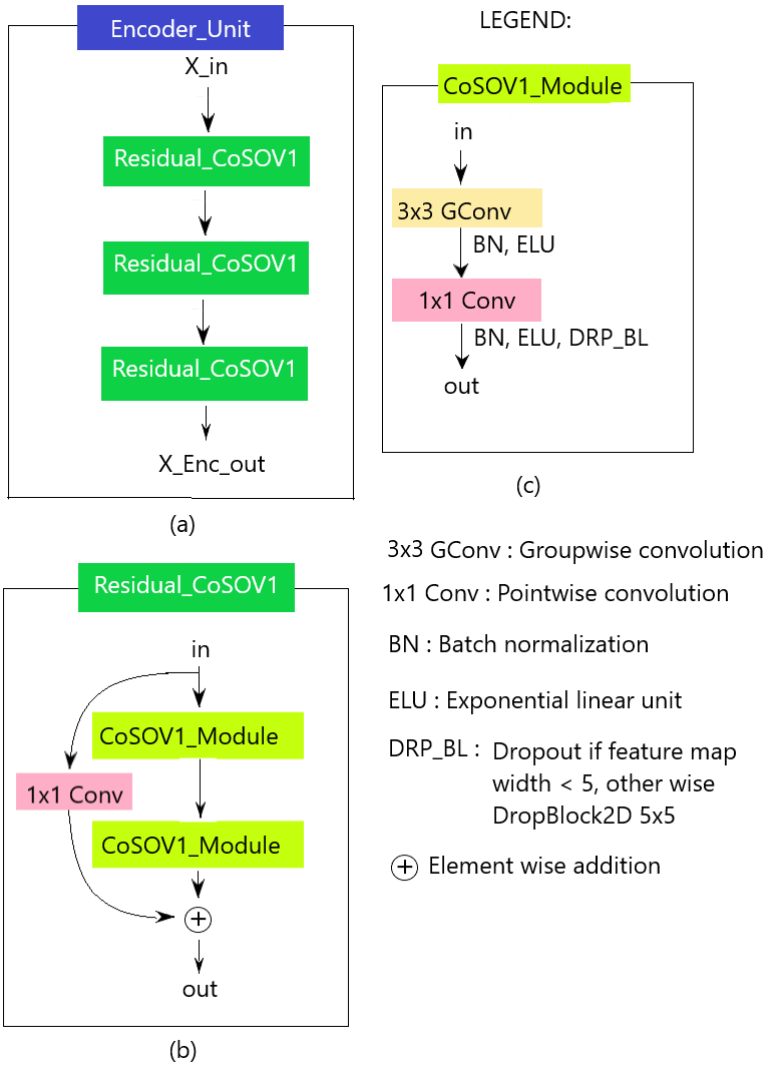

Figure 4. Encoder Unit: (a) Encoder Unit, (b) the residual block , (c) CoSOV1 module.

We used residual block because this kind of block is known to improve training deeper neural networks [48]. The residual block consists of two CoSOV1 modules with a residual link. The reason of all these repetitions is to encode more information and thus allow to increase our network performance.

In the encoder, schematically, as explained above (Section 3.2), the CoSOV1 module (Figure 4 (c)) splits the input channels in groups and apply a group-wise convolution (3×3 convolution). After a point-wise convolution is applied to the outputs of the concatenated groups (see Figure 5 for the first level input illustration). Each of these convolutions is followed by a batch normalization and a non-linear function (ELU: Exponential Linear Unit activation). After these layers, during the model training, the regularization is done in the CoSOV1 module using the dropout [45] method for small feature maps (dimension smaller than 5×5) and the Dropblock [44] for feature maps with dimension greater than 5×5 which is a variant of dropout that zeroes a block instead of pixels individually as Dropout does.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

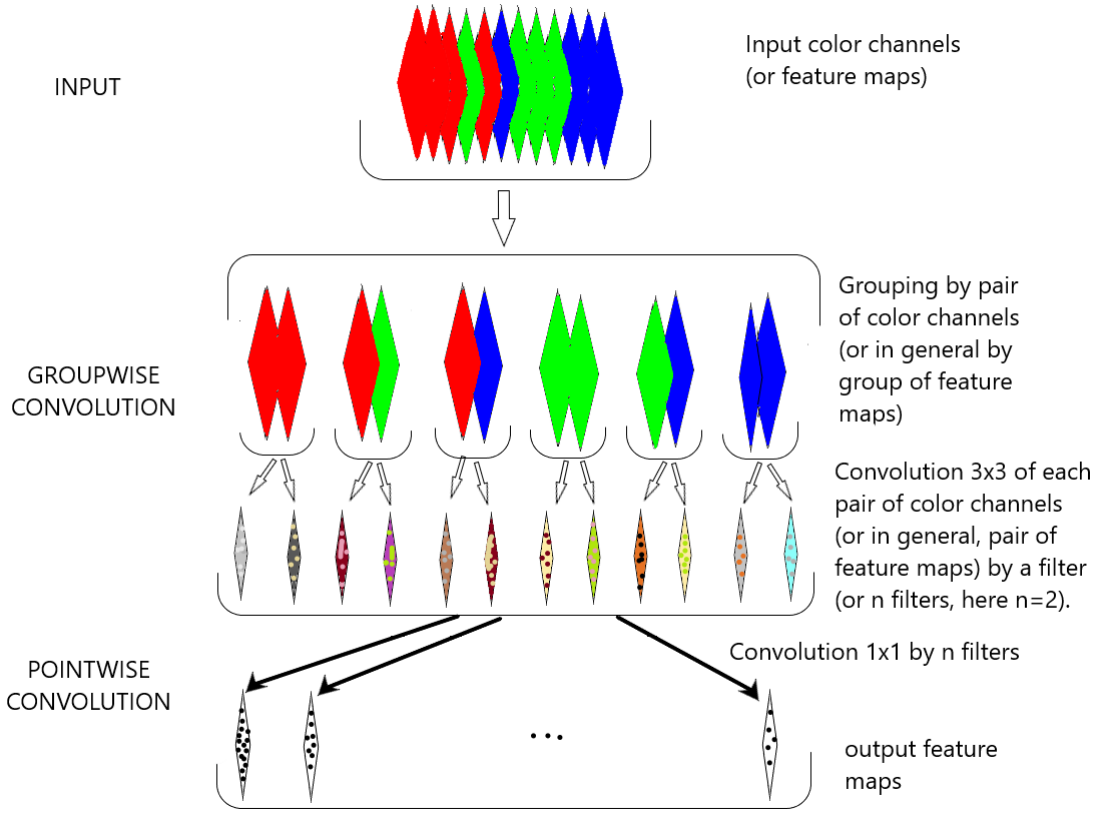

Figure 5. Simplified flowchart in CoSOV1 module for processing pairs of opposing color pairs (or group of feature maps).

At its end, the encoder is followed by the middle unit (see Figure 6 (a)) which is the CoSOV1 module (see Figure 6 (b)) where we remove the group-wise convolution, since at this stage the feature maps are $1 \times 1 \times 128$ in size, and add a residual link.

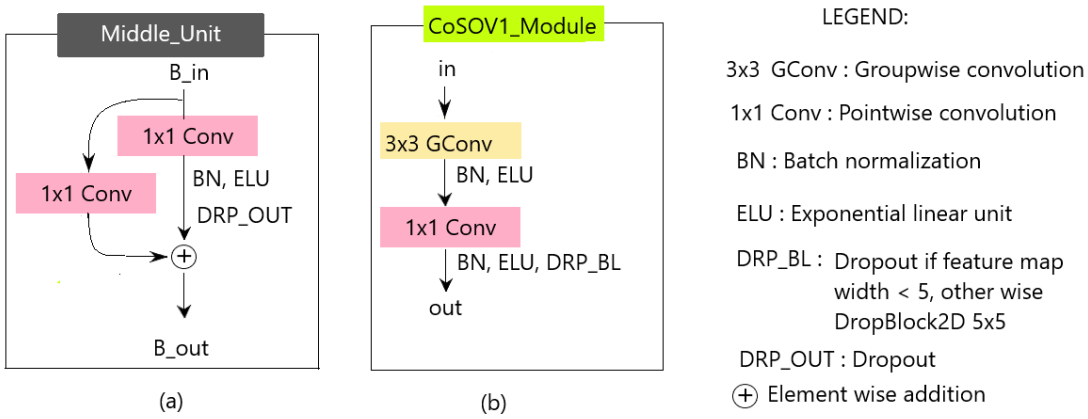

Figure 6. (a) The Middle Unit, (b) the CoSOV1 module.

3.3.3. Decoder

The decoder transforms the features from the encoder to obtain the estimate of the salient object(s) present in the input image. This transformation is achieved through a repeating block, namely the Decoder Unit (see Figure 7 (a)). The Decoder Unit consists of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

two parts: the Decoder Residual Block (see Figure 7 (b)) and the Decoder Deconvolution block (see Figure 7 (c)). The Decoder Residual Block is a modified CoSOV1 module that allows the model to take into account the output of the corresponding level in the encoder. The output of the Decoder Residual Block took two directions. On the one hand, it is passed to the next level of the Decoder and on the other to the second part of the Decoder Unit which is the Decoder Deconvolution block. The latter deconvolves this output obtaining two feature maps having the size of the input image ($384 \times 384 \times 2$ in our case). At the last level of the decoder, all the outputs from the Deconvolution blocks are concatenated and fed to a convolution layer followed by a softmax activation layer which gives the estimation of the salient object detection map.

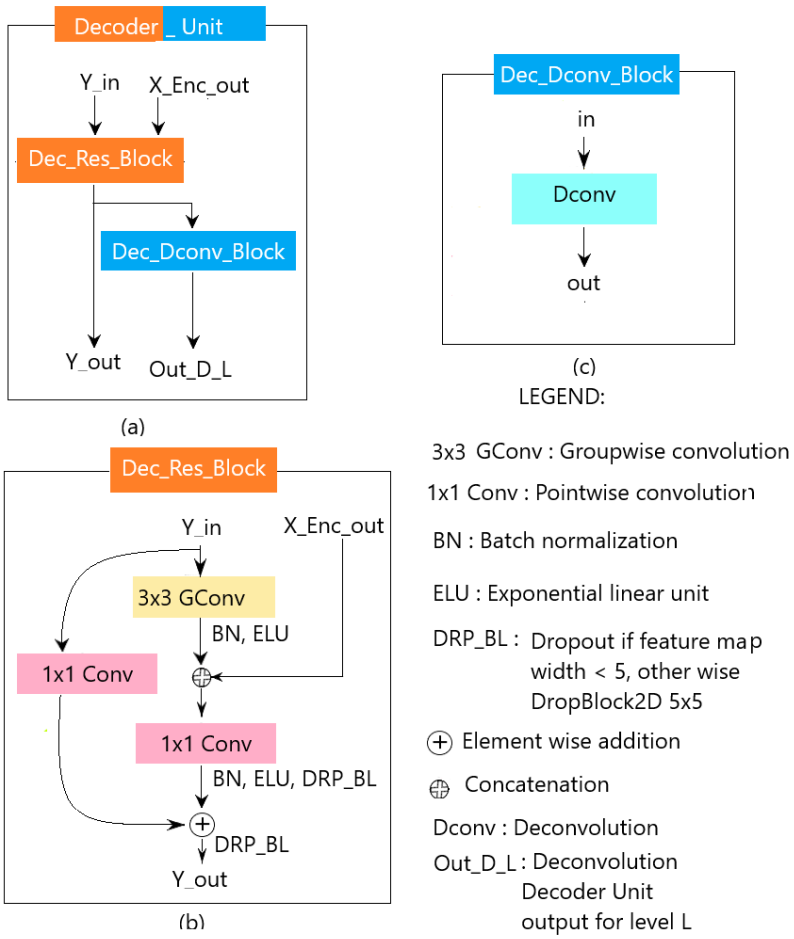

Figure 7. (a) The Decoder Unit; (b) the Decoder Residual Block; (c) the Decoder Deconvolution Block.

4. Experimental Results

4.1. Implementation Details

For our proposed model implementation, we used the deep learning platform TensorFlow with Keras deep learning application programming interface (API) [49]. All input images are resized to 384×384 and pixels values are normalized (each pixel channel value $\in [0.0, \ldots, 1.0]$ and ground truth pixels $\in \{0, 1\}$). Experiments were conducted on a single GPU, nvidia Geforce RTX 3090 TI (24 GB) and an Intel CPU, i7-11700F.

4.2. Datasets

Our proposed model's experiments were conducted on public datasets which are the most widely used in the salient object detection field [50]. Thus, we used Extended Complex

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

Scene Saliency dataset (ECSSD) [51], DUT-OMRON (Dalian University of Technology—OMRON Corporation) [52], DUTS [53], HKU-IS [54], THUR15K [55] datasets.

The ECSSD [51] contains 1000 natural images and their ground truth. Many of its images are semantically meaningful, but structurally complex for saliency detection [51].

DUT-OMRON [52] contains 5168 images and their binary mask with diverse variations and complex background.

DUTS dataset [53] is divided into DUTS-TR (10553 training images) and DUTS-TE (5019 test images). We train and validate our proposed model on the DUTS-TR and DUTS-TE was used for tests.

HKU-IS [54] is composed by 4447 complex images, which contains many disconnected objects with different spatial distributions. Furthermore, it is very challenging for the similar foreground/background appearance [56].

THUR15K is a dataset of images taken from the "Flickr" web site divided into five categories (butterfly, coffee mug, dog jump, giraffe, plane), which contains 3000 images. The images of this dataset represent real world scenes and are considered complex for obtaining salient objects [55] (6232 images with ground truths).

4.3. Model Training Settings

For the reproducibility of the experiments, we set the seed=123. We train our proposed model on DUTS-TR (10553 training images). We split DUTS-TR dataset in a train set (9472 images) and a validation set (1056 images). That is approximately 90% of the dataset for the training set and 10% for the validation set. We didn't use the 25 images because we wanted the training set and the validation set to be divisible by batch size which is 32.

Our proposed model is trained on scratch without pre-trained backbones from images classification (i.e. VGG [57], etc.) or lightweight backbones (i.e. MobileNets [28,29] or ShuffleNets [30,31]). As DUTS-TR is not a big dataset, we used data augmentation during training and many epochs in order to overcome this problem. Indeed, more there are epochs more the data augmentation process transforms data. Thus, our proposed model training has two successive stages:

- The first stage is with data augmentation. The data augmentation is applied on each batch with random transformation (40% zoom in or horizontal flip or vertical flip). This stage has 480 epochs: 240 epochs with learning rate = 0.001 and the following 240 epochs with learning rate=0.0001;
- The second stage is without data augmentation. It has 620 epochs: 240 epochs with learning rate = 0.001, followed by 140 epochs with learning rate = 0.0001 and 240 epochs with learning rate = 0.00005.

We also use a same initializer for all layers in the neural network: HeUniform keras initializer [58] which draws samples from a uniform distribution within [-limit, limit], where limit = $\sqrt{\frac{6}{fan\_in}}$ (fan_in is the number of input units in the weight tensor). The dropout rate is set to 0.2. We used the RMSprop [59] keras optimizer with default values except for the learning rate, the centered which is set to true and the clipnorm=1. The loss function used is the "sparse_categorical_crossentropy" keras function; the keras metrics is "SparseCategoricalAccuracy; the keras check point monitor is "val_sparse_categorical_accuracy"

4.4. Evaluation Metrics

4.4.1. Accuracy

The metrics used to evaluate our proposed model accuracy are: F_β measure, MAE (mean absolute error), weighted F_β^w measure [60]. We also used Precision-Recall and F_β measures curves.

The F_β-measure (F_β) is the weighted harmonic mean of Precision and Recall:

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \quad (4)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

During evaluation $\beta^2 = 0.3$ as it is often suggested [16,56].

Let $\overline{S}$ be the saliency map estimation with pixels values normalized in order to be in $[0.0, \ldots, 1.0]$ and $\overline{G}$, its ground truth also normalized in $\{0; 1\}$. The MAE (mean absolute error) is:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |\overline{S}(x,y) - \overline{G}(x,y)| \tag{5}$$

where W and H are the width and the height respectively of the above maps ($\overline{S}$ and $\overline{G}$).

The F_β^w measure [60] fixes the interpolation flaw, dependence flaw, and equal importance flaw in traditional evaluation metrics and its value is:

$$F_\beta^w = (1 + \beta^2) \frac{\text{Precision}^w \times \text{Recall}^w}{\beta^2 \times \text{Precision}^w + \text{Recall}^w} \tag{6}$$

Precision^w and Recall^w are the weighted *Precision* and the weighted *Recall* respectively.

4.4.2. Lightweight measures

Since, in this work we proposed a lightweight salient object detection model, thus, we also evaluate the model with lightweight measures: the number of parameters, the saliency maps estimation speed (FPS: frames per second) and the computational cost by measuring the FLOPS (the number of floating-point operations). The FLOPS is related to the devices energy consumption (more FLOPS more energy consumption).

4.5. Comparison with state-of-the-art

We compare our proposed model with 20 salient object detection state-of-the-art and 10 lightweight salient object detection state-of-the-art models. We divided these methods because the lightweight methods outperform others with respect to lightweight measures. However, the lightweight methods accuracy is lower than the accuracy of those with huge parameters. We mainly use the salient object detection results provided by Liu *et al.* [16] except for F_β measure and Precision-Recall curves where we use saliency maps provided by these authors. We also used saliency maps provided by the HVPNet authors [17] to compute HVPNet F_β^ω measures.

In this section we describe the comparison with the 20 salient object detection namely DRFI[61], DCL [62], DHSNet [63], RFCN [64], NLDF [65], DSS [66], Amulet [67], UCF [68], SRM [69], PiCANet [70], BRN [71], C2S [72], RAS [73], DNA [74], CPD [75], BASNet [76], AFNet [77], PoolNet [78], EGNet [79] and BANet [80].

Table 1 shows that our proposed model CoSOV1Net outperforms all the 20 salient object detection state-of-the-art for lightweight measures (#parameters, FLOPS and FPS) with a large margin (i.e. the best among them for the FLOPS is DHSNet [63] with $FLOPS = 15.8G$ and $F_\beta = 0.903$ for ECSSD; the worst is EGNet [79] with $FLOPS = 270.8G$ and $F_\beta = 0.938$ for ECSSD; while our proposed model CoSOV1Net has $FLOPS = 1.4G$ and its $F_\beta = 0.931$ for ECSSD) (see Table 1).

Table 1 also shows that CoSOV1Net is among the top 6 models for ECSSD, among the top 7 for DUT-OMRON around the top 10 for the other three datasets for F-measure. Table 2 and Table 3 compare our model with the state-of-the-art models for respectively MAE and F_β^ω measures. From this comparison, we see that our model is ranked around the top 10 for all the 4 datasets and the 15^{th} rank for the HKU-IS dataset. This demonstrates that our model is also competitive with respect to the performance of state-of-the-art models.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

Table 1. Our proposed model F-measure ($F_\beta \uparrow, \beta^2 = 0.3$) compared with 20 state-of-the-art models (Best value in bold).

Methods	#Param (M)↓	FLOPS (G) ↓	Speed (FPS) ↑	ECSSD	DUT-OMRON	DUTS-TE	HKU-IS	THUR15K
DRFI[61]	-	-	0.1	0.777	0.652	0.649	0.774	0.670
DCL [62]	66.24	224.9	1.4	0.895	0.733	0.785	0.892	0.747
DHSNet [63]	94.04	15.8	10.0	0.903	-	0.807	0.889	0.752
RFCN [64]	134.69	102.8	0.4	0.896	0.738	0.782	0.892	0.754
NLDF [65]	35.49	263.9	18.5	0.902	0.753	0.806	0.902	0.762
DSS [66]	62.23	114.6	7.0	0.915	0.774	0.827	0.913	0.770
Amulet [67]	33.15	45.3	9.7	0.913	0.743	0.778	0.897	0.755
UCF [68]	23.98	61.4	12.0	0.901	0.730	0.772	0.888	0.758
SRM [69]	43.74	20.3	12.3	0.914	0.769	0.826	0.906	0.778
PiCANet [70]	32.85	37.1	5.6	0.923	0.766	0.837	0.916	0.783
BRN [71]	126.35	24.1	3.6	0.919	0.774	0.827	0.910	0.769
C2S [72]	137.03	20.5	16.7	0.907	0.759	0.811	0.898	0.775
RAS [73]	20.13	35.6	20.4	0.916	0.785	0.831	0.913	0.772
DNA [74]	20.06	82.5	25.0	0.935	0.799	0.865	0.930	0.793
CPD [75]	29.23	59.5	68.0	0.930	0.794	0.861	0.924	0.795
BASNet [76]	87.06	127.3	36.2	0.938	**0.805**	0.859	0.928	0.783
AFNet [77]	37.11	38.4	21.6	0.930	0.784	0.857	0.921	0.791
PoolNet [78]	53.63	123.4	39.7	0.934	0.791	0.866	0.925	**0.800**
EGNet [79]	108.07	270.8	12.7	0.938	0.794	0.870	0.928	**0.800**
BANet [80]	55.90	121.6	12.5	**0.940**	0.803	**0.872**	**0.932**	0.796
CoSOV1Net (OURS)	**1.14**	**1.4**	**211.2**	0.931	0.789	0.833	0.912	0.773

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

Table 2. Our proposed model MAE (↓) compared with 20 state-of-the-art models (Best performance in bold).

Methods	#Param (M)↓	FLOPS (G) ↓	Speed (FPS) ↑	ECSSD	DUT-OMRON	DUTS-TE	HKU-IS	THUR15K
DRFI[61]	-	-	0.1	0.161	0.138	0.154	0.146	0.150
DCL [62]	66.24	224.9	1.4	0.080	0.095	0.082	0.063	0.096
DHSNet [63]	94.04	15.8	10.0	0.062	-	0.066	0.053	0.082
RFCN [64]	134.69	102.8	0.4	0.097	0.095	0.089	0.080	0.100
NLDF [65]	35.49	263.9	18.5	0.066	0.080	0.065	0.048	0.080
DSS [66]	62.23	114.6	7.0	0.056	0.066	0.056	0.041	0.074
Amulet [67]	33.15	45.3	9.7	0.061	0.098	0.085	0.051	0.094
UCF [68]	23.98	61.4	12.0	0.071	0.120	0.112	0.062	0.112
SRM [69]	43.74	20.3	12.3	0.056	0.069	0.059	0.046	0.077
PiCANet [70]	32.85	37.1	5.6	0.049	0.068	0.054	0.042	0.083
BRN [71]	126.35	24.1	3.6	0.043	0.062	0.050	0.036	0.076
C2S [72]	137.03	20.5	16.7	0.057	0.072	0.062	0.046	0.083
RAS [73]	20.13	35.6	20.4	0.058	0.063	0.059	0.045	0.075
DNA [74]	20.06	82.5	25.0	0.041	**0.056**	0.044	**0.031**	0.069
CPD [75]	29.23	59.5	68.0	0.044	0.057	0.043	0.033	**0.068**
BASNet [76]	87.06	127.3	36.2	0.040	**0.056**	0.048	0.032	0.073
AFNet [77]	37.11	38.4	21.6	0.045	0.057	0.046	0.036	0.072
PoolNet [78]	53.63	123.4	39.7	0.048	0.057	0.043	0.037	**0.068**
EGNet [79]	108.07	270.8	12.7	0.044	**0.056**	0.044	0.034	0.070
BANet [80]	55.90	121.6	12.5	**0.038**	0.059	**0.040**	**0.031**	**0.068**
CoSOV1Net (OURS)	**1.14**	**1.4**	**211.2**	0.051	0.064	0.057	0.045	0.076

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

Table 3. Our proposed model Weighted F-measure ($F_\beta^\omega \uparrow, \beta^2 = 1$) compared with 20 state-of-the-art models (Best value in bold).

Methods	#Param (M)↓	FLOPS (G) ↓	Speed (FPS) ↑	ECSSD	DUT-OMRON	DUTS-TE	HKU-IS	THUR15K
DRFI[61]	-	-	0.1	0.548	0.424	0.378	0.504	0.444
DCL [62]	66.24	224.9	1.4	0.782	0.584	0.632	0.770	0.624
DHSNet [63]	94.04	15.8	10.0	0.837	-	0.705	0.816	0.666
RFCN [64]	134.69	102.8	0.4	0.725	0.562	0.586	0.707	0.591
NLDF [65]	35.49	263.9	18.5	0.835	0.634	0.710	0.838	0.676
DSS [66]	62.23	114.6	7.0	0.864	0.688	0.752	0.862	0.702
Amulet [67]	33.15	45.3	9.7	0.839	0.626	0.657	0.817	0.650
UCF [68]	23.98	61.4	12.0	0.805	0.573	0.595	0.779	0.613
SRM [69]	43.74	20.3	12.3	0.849	0.658	0.721	0.835	0.684
PiCANet [70]	32.85	37.1	5.6	0.862	0.691	0.745	0.847	0.687
BRN [71]	126.35	24.1	3.6	0.887	0.709	0.774	0.875	0.712
C2S [72]	137.03	20.5	16.7	0.849	0.663	0.717	0.835	0.685
RAS [73]	20.13	35.6	20.4	0.855	0.695	0.739	0.849	0.691
DNA [74]	20.06	82.5	25.0	0.897	0.729	0.797	**0.889**	0.723
CPD [75]	29.23	59.5	68.0	0.889	0.715	0.799	0.879	**0.731**
BASNet [76]	87.06	127.3	36.2	0.898	**0.751**	0.802	**0.889**	0.721
AFNet [77]	37.11	38.4	21.6	0.880	0.717	0.784	0.869	0.719
PoolNet [78]	53.63	123.4	39.7	0.875	0.710	0.783	0.864	0.724
EGNet [79]	108.07	270.8	12.7	0.886	0.727	0.796	0.876	0.727
BANet [80]	55.90	121.6	12.5	**0.901**	0.736	**0.810**	**0.889**	0.730
CoSOV1Net (OURS)	**1.14**	**1.4**	**211.2**	0.861	0.696	0.731	0.834	0.688

We also compare our proposed model CoSOV1Net with the lightweight salient object detection state-of-the-art: MobileNet [28], MobileNetV2 [29], ShuffleNet [30], ShuffleNetV2 [31], ICNet [81], BiSeNet R18 [82], BiSeNet X39 [82], DFANet [83], HVPNet [17], SAMNet [16].

For the comparison with lightweight state-of-the-art, Table 4 shows that our proposed model outperforms these lightweight state-of-the-art for the models parameters number and F_β measure for ECSSD dataset and is competitive for other measure and dataset. Table 5 shows that our model outperforms these lightweight state-of-the-art for MAE measure for ECSSD and DUTS-TE datasets, is ranked 1^{st} ex aequo with HVPNet for DUT-OMRON, ranked 1^{st} ex aequo with HVPNet and SAMNet for HKU-IS dataset and 2^{nd} for THUR15K dataset. Our model also outperforms these lightweight state-of-the-art for F_β^ω measure for ECSSD and DUTS-TE and is competitive for the 3 other datasets (see Table 6).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

Table 4. Our proposed model F-measure ($F_\beta \uparrow$, $\beta^2 = 0.3$) compared with lightweight salient object detection state-of-the-art models (Best value in bold).

Methods	#Param (M)↓	FLOPS (G) ↓	Speed (FPS) ↑	ECSSD	DUT-OMRON	DUTS-TE	HKU-IS	THUR15K
MobileNet* [28]	4.27	2.2	295.8	0.906	0.753	0.804	0.895	0.767
MobileNetV2* [29]	2.37	0.8	446.2	0.905	0.758	0.798	0.890	0.766
ShuffleNet* [30]	1.80	0.7	406.9	0.907	0.757	0.811	0.898	0.771
ShuffleNetV2* [31]	1.60	**0.5**	**452.5**	0.901	0.746	0.789	0.884	0.755
ICNet [81]	6.70	6.3	75.1	0.918	0.773	0.810	0.898	0.768
BiSeNet R18 [82]	13.48	25.0	120.5	0.909	0.757	0.815	0.902	0.776
BiSeNet X39 [82]	1.84	7.3	165.8	0.901	0.755	0.787	0.888	0.756
DFANet [83]	1.83	1.7	91.4	0.896	0.750	0.791	0.884	0.757
HVPNet [17]	1.23	1.1	333.2	0.925	**0.799**	**0.839**	**0.915**	**0.787**
SAMNet [16]	1.33	**0.5**	343.2	0.925	0.797	0.835	**0.915**	0.785
CoSOV1Net (OURS)	**1.14**	1.4	211.2	**0.931**	0.789	0.833	0.912	0.773

* SAMNet where the encoder is replaced by this backbone.

Table 5. Our proposed model MAE (↓) compared with lightweight salient object detection state-of-the art models (Best value in bold).

Methods	#Param (M)↓	FLOPS (G) ↓	Speed (FPS) ↑	ECSSD	DUT-OMRON	DUTS-TE	HKU-IS	THUR15K
MobileNet* [28]	4.27	2.2	295.8	0.064	0.073	0.066	0.052	0.081
MobileNetV2* [29]	2.37	0.8	446.2	0.066	0.075	0.070	0.056	0.085
ShuffleNet* [30]	1.80	0.7	406.9	0.062	0.069	0.062	0.050	0.078
ShuffleNetV2* [31]	1.60	**0.5**	**452.5**	0.069	0.076	0.071	0.059	0.086
ICNet [81]	6.70	6.3	75.1	0.059	0.072	0.067	0.052	0.084
BiSeNet R18 [82]	13.48	25.0	120.5	0.062	0.072	0.062	0.049	0.080
BiSeNet X39 [82]	1.84	7.3	165.8	0.070	0.078	0.074	0.059	0.090
DFANet [83]	1.83	1.7	91.4	0.073	0.078	0.075	0.061	0.089
HVPNet [17]	1.23	1.1	333.2	0.055	**0.064**	0.058	**0.045**	0.076
SAMNet [16]	1.33	**0.5**	343.2	0.053	0.065	0.058	**0.045**	**0.077**
CoSOV1Net (OURS)	**1.14**	1.4	211.2	**0.051**	**0.064**	**0.057**	**0.045**	0.076

* SAMNet where the encoder is replaced by this backbone.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

Table 6. Our proposed model Weighted F-measure ($F_\beta^\omega \uparrow$, $\beta^2 = 1$) compared with lightweight salient object detection models (Best value in bold).

Methods	#Param (M)↓	FLOPS (G) ↓	Speed (FPS) ↑	ECSSD	DUT-OMRON	DUTS-TE	HKU-IS	THUR15K
MobileNet* [28]	4.27	2.2	295.8	0.829	0.656	0.696	0.816	0.675
MobileNetV2* [29]	2.37	0.8	446.2	0.820	0.651	0.676	0.799	0.660
ShuffleNet* [30]	1.80	0.7	406.9	0.831	0.667	0.709	0.820	0.683
ShuffleNetV2* [31]	1.60	**0.5**	**452.5**	0.812	0.637	0.665	0.788	0.652
ICNet [81]	6.70	6.3	75.1	0.838	0.669	0.694	0.812	0.668
BiSeNet R18 [82]	13.48	25.0	120.5	0.829	0.648	0.699	0.819	0.675
BiSeNet X39 [82]	1.84	7.3	165.8	0.802	0.632	0.652	0.784	0.641
DFANet [83]	1.83	1.7	91.4	0.799	0.627	0.652	0.778	0.639
HVPNet [17]	1.23	1.1	333.2	0.854	**0.699**	0.730	**0.839**	**0.696**
SAMNet [16]	1.33	**0.5**	343.2	0.855	**0.699**	0.729	0.837	0.693
CoSOV1Net (OURS)	**1.14**	1.4	211.2	**0.861**	0.696	**0.731**	0.834	0.688

* SAMNet where the encoder is replaced by this backbone.

4.6. Comparison with SAMNet and HVPNet state-of-the-art

We chose to compare our CoSOV1Net model specifically with SAMNet [16] and HVPNet [17] because they are among the best state-of-the-art models. Figure 8 and Figure 9 show that the proposed model is competitive with these two lightweight salient object detection state-of-the-art with respect to precision-recall and F_β measure curves.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

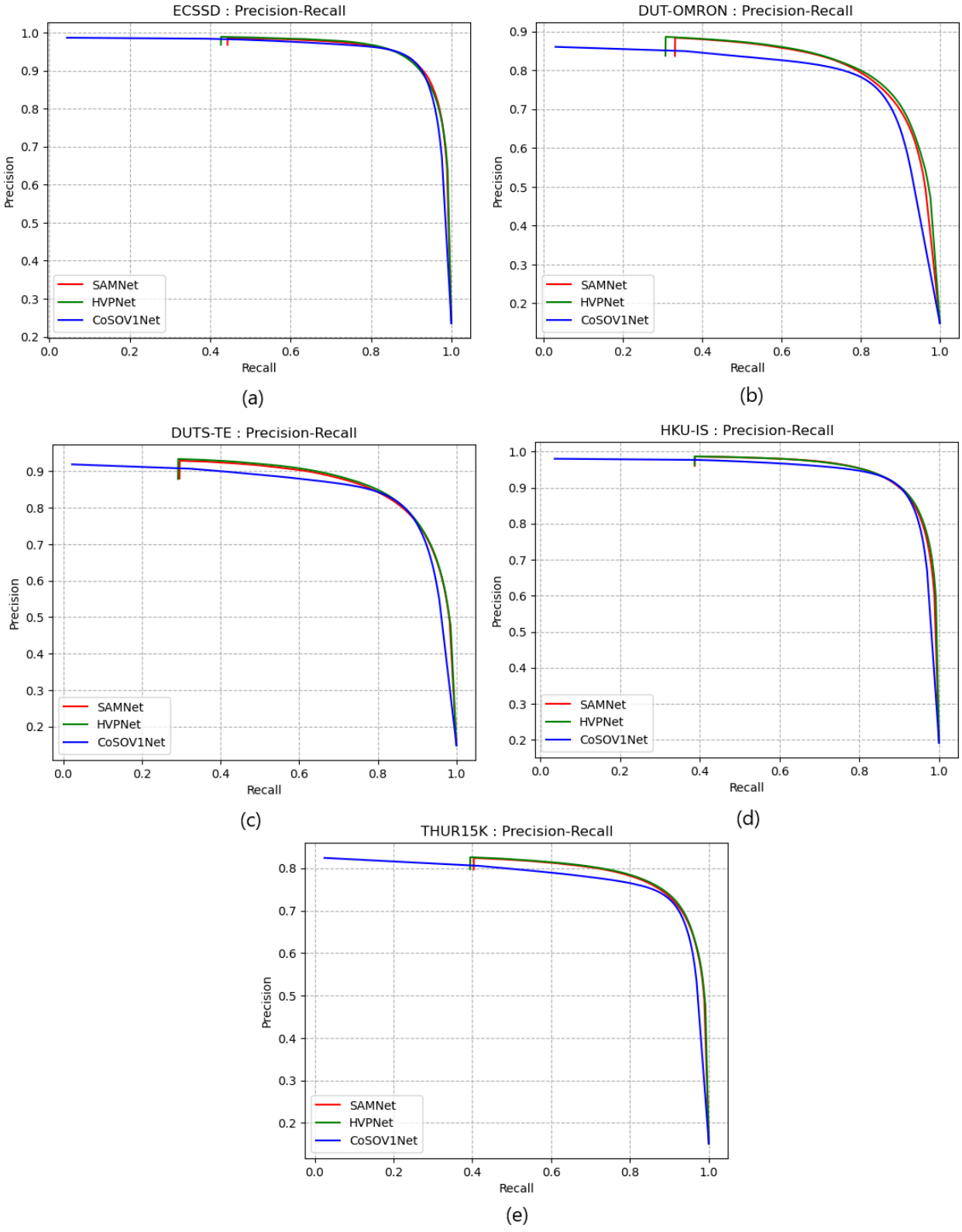

Figure 8. Precision Recall curves for (a) ECSSD, (b) DUT-OMRON, (c) DUTS-TE, (d) HKU-IS and (e) THUR15K datasets.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

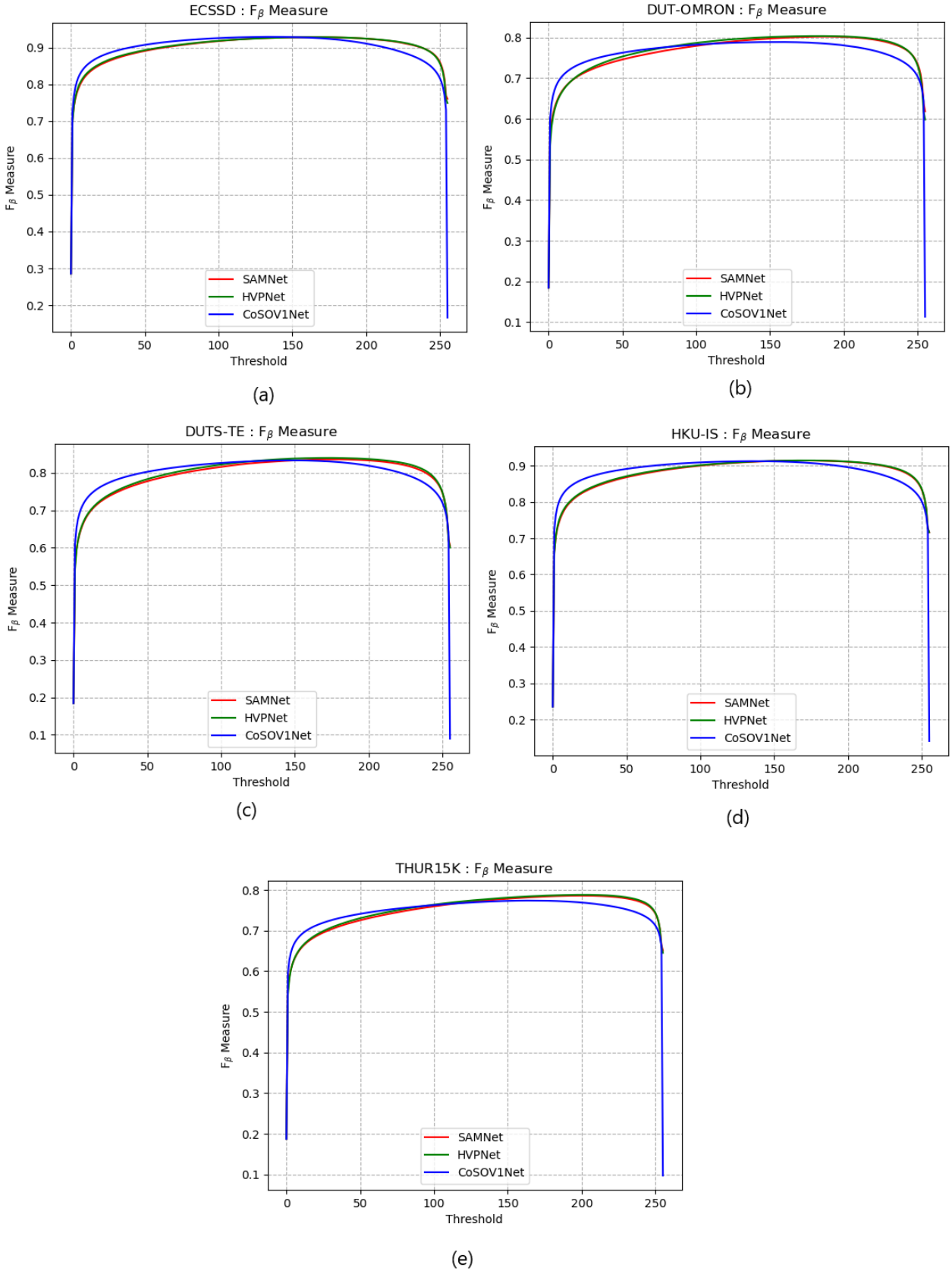

Figure 9. F_β measure curves for (a) ECSSD, (b) DUT-OMRON, (c) DUTS-TE, (d) HKU-IS and (e) THUR15K datasets.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

For qualitative comparison, Figure 10 shows some images highlighting that our proposed model (CoSOV1Net) is competitive with regard to the state-of-the-art SAMNet [16] and HVPNet [17] models which are among the best ones.

Images from columns 1 and 2 show a big salient object on a cloudy background, and a big object on a complex background respectively: CoSOV1Net (ours) performs better than HVPNet on these saliency maps. Columns 3 shows salient objects with same colors and column 4 shows salient objects with multiple colors: SAMNet and CoSOV1Net saliency maps are slightly identical and HVPNet saliency map is slightly better. Column 5 shows image with 3 salient objects with different sizes and colors: 2 are big and 1 is very small; CoSOV1Net saliency map is better than the SAMNet's and the HVPNet's. Column 6 shows red salient objects on black and yellow background; SAMNet's saliency map is worst while CoSOV1Net and HVPNet perform well on that image. Column 7 shows a complex background and multiple salient objects with different colors: CoSOV1Net performs better than SAMNet and HVPNet. Column 8 shows tiny salient objects: the 3 models perform well. On column 9, SAMNet has worst performance while CoSOV1Net is the best. Column 10 shows colored glasses as salient objects: the CoSOV1Net performance is better than the SAMNet's and HVPNet's. On column 11, SAMNet has worst performance. On column 12 and 13, CoSOV1Net has the best performance. Column 18 shows a sub-marine image: CoSOV1Net is better than SAMNet.

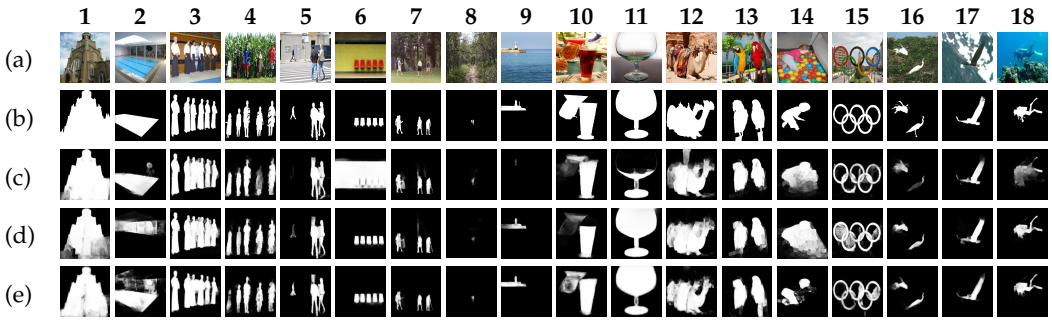

Figure 10. Comparison between SAMNet [16], HVPNet [17] and our proposed model CoSOV1Net on some images saliency maps:(a) Images; (b) Ground Truth; (c) SAMNet; (d) HVPNet; (e) CoSOV1Net (Ours).

5. Discussion

The results shows the performance of our model CoSOV1Net for accuracy measures and lightweight measures. The CoSOV1Net's rank, when compared to state-of-the-art models, shows that CoSOV1Net behaves as a lightweight salient object detection by dominating the lightweight measures and having good performance for accuracy measures (see Table 7).

Table 7. Our proposed model (CoSOV1Net) ranking with respect to existing salient object detection.

Measure	#Param (M)↓	FLOPS (G) ↓	Speed (FPS) ↑	ECSSD	DUT-OMRON	DUTS-TE	HKU-IS	THUR15K
F_β	1^{st}	1^{st}	1^{st}	6^{th}	7^{th}	9^{th}	11^{th}	11^{th}
MAE	1^{st}	1^{st}	1^{st}	10^{th}	10^{th}	11^{th}	11^{th}	10^{th}
F_β^ω	1^{st}	1^{st}	1^{st}	11^{th}	9^{th}	11^{th}	15^{th}	11^{th}

The results also showed that when the proposed model CoSOV1Net is compared to the lightweight salient object detection state-of-the-art, its measures results are ranked generally among the best for datasets and measures used. Thus, we can conclude that CoSOV1Net behaves as a competitive lightweight salient object detection model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

Table 8. Our proposed model (CoSOV1Net) ranking with respect to lightweight salient object detection models.

Measure	#Param (M)↓	FLOPS (G) ↓	Speed (FPS) ↑	ECSSD	DUT-OMRON	DUTS-TE	HKU-IS	THUR15K
F_β	1^{st}	6^{th}	7^{th}	1^{st}	3^{rd}	3^{rd}	3^{rd}	4^{th}
MAE	1^{st}	6^{th}	7^{th}	1^{st}	1^{st}	1^{st}	1^{st}	2^{nd}
F_β^ω	1^{st}	6^{th}	7^{th}	1^{st}	3^{rd}	1^{st}	3^{rd}	3^{rd}

As we didn't use backbones from images classification (i.e. VGG [57],. . .) or lightweight backbones (i.e. MobileNets [28,29] or ShuffleNets [30,31]), we conclude that our proposed model CoSOV1Net performance is intrinsic to this model itself.

Finally, putting together the measures for the salient object detection models and the lightweight salient object detection models in a graphic, we notice that the CoSOV1Net model is located for F_β measures with respect to FLOPS and for the number of parameters up left while for the FPS measure it is located up right thus showing its performance as lightweight salient object detection model (see Figure 11). This shows that our proposed model CoSOV1Net is competitive with the best state-of-the-art models used.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2023 doi:10.20944/preprints202306.0899.v1

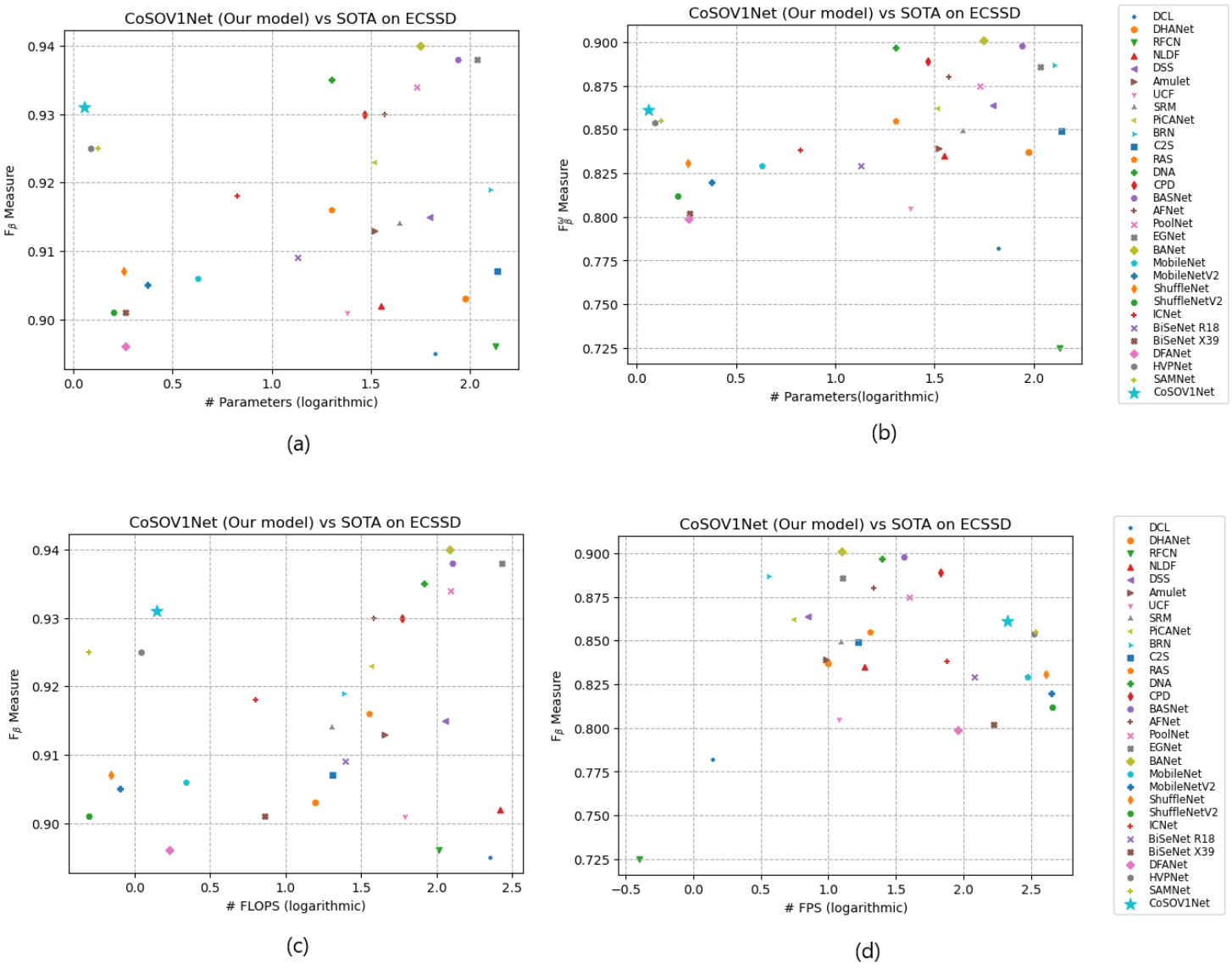

Figure 11. Example of trade-off between (a) F_β measure and #parameters, (b) F_β^ω measure and #parameters, (c) F_β measure and FLOPS, (d)F_β measure and FPS, for ECSSD.

The quantitative and the qualitative comparisons with SAMNet [16] and HVPNet [17] showed that our proposed model has good performance, given these state-of-the-art models are among the best ones.

6. Conclusion

In this work, we presented a lightweight salient object detection deep neural network, CoSOV1Net with very low parameters number (1.14M), low floating-point operations number (FLOPS=1.4G) thus low computational cost and respectable speed ($FPS = 211.2$ on GPU: nvidia Geforce RTX 3090 TI) yet with comparable performance with state-of-the-art salient object detection models that use significantly more parameters and other lightweight salient object detection such as SAMNet [16] and HVPNet [17].

The novelty of our proposed model (CoSOV1Net) is that it uses the principle of integrating color in pattern in a salient object detection deep neural network, since according to Shapley [25] and Shapley *et* Hawken [18] color and pattern are inextricably linked in color human perception. This is implemented by taking inspiration from the primary visual

cortex (V1) cells especially cone- and spatial-opponent cells. Thus, our method extracts features at the color channels spatial level and between the color channels at the same time on a pair of opposing color channels. The idea of grouping color pushed us to group feature maps through the neural network and extract features at the spatial level and between feature maps as done for color channels.

Our results showed that this strategy generates a model which is very promising, competitive with most salient object detection state-of-the-art and lightweight salient object detection state-of-the-art, and practical for mobile environments and limited resources devices.

As future work, our proposed CoSOV1Net model, based on integrating color into patterns, can be improved by coupling it with human visual system attention mechanism, which is the basis of many lightweight models, to produce a more efficient lightweight salient object detection model.

Author Contributions:

"Conceptualization, D.N. and M.M.; methodology, D.N.; software, D.N.; validation, D.N.; formal analysis, D.N.; investigation, D.N.; resources, D.N. and M.M.; data curation, D.N.; writing—original draft preparation, D.N.; writing—review and editing, D.N. and M.M.; visualization, D.N.; supervision, M.M.; project administration, M.M.; funding acquisition, M.M. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: This research was funded by individual discovery grant number RGPIN-2022-03654.

Acknowledgments: Authors would like to thank the NSERC (Natural Sciences and Engineering Research Council of Canada) for having supported this research work via the individual discovery grant program (RGPIN-2022-03654).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ndayikengurukiye, D.; Mignotte, M. Salient Object Detection by LTP Texture Characterization on Opposing Color Pairs under SLICO Superpixel Constraint. *Journal of Imaging* **2022**, *8*, 110.
2. Smeulders, A.W.; Chu, D.M.; Cucchiara, R.; Calderara, S.; Dehghan, A.; Shah, M. Visual tracking: An experimental survey. *IEEE transactions on pattern analysis and machine intelligence* **2013**, *36*, 1442–1468.
3. Pieters, R.; Wedel, M. Attention capture and transfer in advertising: Brand, pictorial, and text-size effects. *Journal of marketing* **2004**, *68*, 36–50.
4. Itti, L. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE transactions on image processing* **2004**, *13*, 1304–1318.
5. Li, J.; Feng, X.; Fan, H. Saliency-based image correction for colorblind patients. *Computational Visual Media* **2020**, *6*, 169–189.
6. Pinciroli Vago, N.O.; Milani, F.; Fraternali, P.; da Silva Torres, R. Comparing CAM algorithms for the identification of salient image features in iconography artwork analysis. *Journal of Imaging* **2021**, *7*, 106.
7. Gao, Y.; Shi, M.; Tao, D.; Xu, C. Database saliency for fast image retrieval. *IEEE Transactions on Multimedia* **2015**, *17*, 359–369.
8. Wong, L.K.; Low, K.L. Saliency-enhanced image aesthetics class prediction. In Proceedings of the 2009 16th IEEE international conference on image processing (ICIP). IEEE, 2009, pp. 997–1000.
9. Liu, H.; Heynderickx, I. Studying the added value of visual attention in objective image quality metrics based on eye movement data. In Proceedings of the 2009 16th IEEE international conference on image processing (ICIP). IEEE, 2009, pp. 3097–3100.
10. Chen, L.Q.; Xie, X.; Fan, X.; Ma, W.Y.; Zhang, H.J.; Zhou, H.Q. A visual attention model for adapting images on small displays. *Multimedia systems* **2003**, *9*, 353–364.
11. Chen, T.; Cheng, M.M.; Tan, P.; Shamir, A.; Hu, S.M. Sketch2photo: Internet image montage. *ACM transactions on graphics (TOG)* **2009**, *28*, 1–10.

12. Huang, H.; Zhang, L.; Zhang, H.C. Arcimboldo-like collage using internet images. In Proceedings of the Proceedings of the 2011 SIGGRAPH Asia Conference, 2011, pp. 1–8.
13. Gupta, A.K.; Seal, A.; Prasad, M.; Khanna, P. Salient object detection techniques in computer vision—A survey. *Entropy* **2020**, *22*, 1174.
14. Wang, W.; Lai, Q.; Fu, H.; Shen, J.; Ling, H.; Yang, R. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2021**, *44*, 3239–3259.
15. Gao, S.H.; Tan, Y.Q.; Cheng, M.M.; Lu, C.; Chen, Y.; Yan, S. Highly efficient salient object detection with 100k parameters. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI. Springer, 2020, pp. 702–721.
16. Liu, Y.; Zhang, X.Y.; Bian, J.W.; Zhang, L.; Cheng, M.M. SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection. *IEEE Transactions on Image Processing* **2021**, *30*, 3804–3814.
17. Liu, Y.; Gu, Y.C.; Zhang, X.Y.; Wang, W.; Cheng, M.M. Lightweight salient object detection via hierarchical visual perception learning. *IEEE Transactions on Cybernetics* **2020**, *51*, 4439–4449.
18. Shapley, R.; Hawken, M.J. Color in the cortex: single-and double-opponent cells. *Vision research* **2011**, *51*, 701–717.
19. Kruger, N.; Janssen, P.; Kalkan, S.; Lappe, M.; Leonardis, A.; Piater, J.; Rodriguez-Sanchez, A.J.; Wiskott, L. Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE transactions on pattern analysis and machine intelligence* **2012**, *35*, 1847–1871.
20. Nunez, V.; Shapley, R.M.; Gordon, J. Cortical double-opponent cells in color perception: perceptual scaling and chromatic visual evoked potentials. *i-Perception* **2018**, *9*, 2041669517752715.
21. Conway, B.R. Color vision, cones, and color-coding in the cortex. *The neuroscientist* **2009**, *15*, 274–290.
22. Conway, B.R. Spatial structure of cone inputs to color cells in alert macaque primary visual cortex (V-1). *Journal of Neuroscience* **2001**, *21*, 2768–2783.
23. Hunt, R.W.G.; Pointer, M.R. *Measuring colour*; John Wiley & Sons, 2011.
24. Engel, S.; Zhang, X.; Wandell, B. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature* **1997**, *388*, 68–71.
25. Shapley, R. Physiology of color vision in primates. In *Oxford Research Encyclopedia of Neuroscience*; 2019.
26. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern recognition* **2020**, *106*, 107404.
27. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, 2015, pp. 234–241.
28. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* **2017**.
29. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.
30. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6848–6856.
31. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 116–131.
32. Frintrop, S.; Werner, T.; Martin Garcia, G. Traditional saliency reloaded: A good old model in new shape. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 82–90.
33. Mäenpää, T.; Pietikäinen, M. Classification with color and texture: jointly or separately? *Pattern recognition* **2004**, *37*, 1629–1640.
34. Chan, C.H.; Kittler, J.; Messer, K. Multispectral local binary pattern histogram for component-based color face verification. In Proceedings of the 2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems. IEEE, 2007, pp. 1–7.

35. Faloutsos, C.; Lin, K.I. *FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets*; Vol. 24, ACM, 1995.
36. Jain, A.; Healey, G. A multiscale representation including opponent color features for texture recognition. *IEEE Transactions on Image Processing* **1998**, *7*, 124–128.
37. Yang, K.F.; Gao, S.B.; Guo, C.F.; Li, C.Y.; Li, Y.J. Boundary detection using double-opponency and spatial sparseness constraint. *IEEE Transactions on Image Processing* **2015**, *24*, 2565–2578.
38. Hurvich, L.M.; Jameson, D. An opponent-process theory of color vision. *Psychological review* **1957**, *64*, 384.
39. Farabet, C.; Couprie, C.; Najman, L.; LeCun, Y. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence* **2012**, *35*, 1915–1929.
40. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International conference on machine learning. pmlr, 2015, pp. 448–456.
41. Santurkar, S.; Tsipras, D.; Ilyas, A.; Madry, A. How does batch normalization help optimization? *Advances in neural information processing systems* **2018**, *31*.
42. Chollet, F. *Deep learning with Python*; Simon and Schuster, 2021.
43. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.
44. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems* **2018**, *31*.
45. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* **2014**, *15*, 1929–1958.
46. Pietikäinen, M.; Hadid, A.; Zhao, G.; Ahonen, T. *Computer vision using local binary patterns*; Vol. 40, Springer Science & Business Media, 2011.
47. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience* **2018**, *2018*.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
49. Chollet, F.; et al. Keras. https://keras.io, 2015.
50. Borji, A.; Cheng, M.M.; Jiang, H.; Li, J. Salient object detection: A benchmark. *IEEE transactions on image processing* **2015**, *24*, 5706–5722.
51. Shi, J.; Yan, Q.; Xu, L.; Jia, J. Hierarchical image saliency detection on extended CSSD. *IEEE transactions on pattern analysis and machine intelligence* **2016**, *38*, 717–729.
52. Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; Yang, M.H. Saliency detection via graph-based manifold ranking. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 3166–3173.
53. Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; Ruan, X. Learning to detect salient objects with image-level supervision. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 136–145.
54. Li, G.; Yu, Y. Visual saliency based on multiscale deep features. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5455–5463.
55. Cheng, M.M.; Mitra, N.J.; Huang, X.; Hu, S.M. Salientshape: group saliency in image collections. *The visual computer* **2014**, *30*, 443–453.
56. Feng, W.; Li, X.; Gao, G.; Chen, X.; Liu, Q. Multi-scale global contrast CNN for salient object detection. *Sensors* **2020**, *20*, 2656.
57. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* **2014**.
58. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2015, pp. 1026–1034.
59. Tieleman, T.; Hinton, G.; et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* **2012**, *4*, 26–31.
60. Margolin, R.; Zelnik-Manor, L.; Tal, A. How to evaluate foreground maps? In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 248–255.

61. Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; Li, S. Salient object detection: A discriminative regional feature integration approach. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2083–2090.
62. Li, G.; Yu, Y. Deep contrast learning for salient object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 478–487.
63. Liu, N.; Han, J. Dhsnet: Deep hierarchical saliency network for salient object detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 678–686.
64. Wei, J.; Zhong, B. Saliency detection using fully convolutional network. In Proceedings of the 2018 Chinese Automation Congress (CAC). IEEE, 2018, pp. 3902–3907.
65. Luo, Z.; Mishra, A.; Achkar, A.; Eichel, J.; Li, S.; Jodoin, P.M. Non-local deep features for salient object detection. In Proceedings of the Proceedings of the IEEE Conference on computer vision and pattern recognition, 2017, pp. 6609–6617.
66. Hou, Q.; Cheng, M.M.; Hu, X.; Borji, A.; Tu, Z.; Torr, P.H. Deeply supervised salient object detection with short connections. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3203–3212.
67. Zhang, P.; Wang, D.; Lu, H.; Wang, H.; Ruan, X. Amulet: Aggregating multi-level convolutional features for salient object detection. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 202–211.
68. Zhang, P.; Wang, D.; Lu, H.; Wang, H.; Yin, B. Learning uncertain convolutional features for accurate saliency detection. In Proceedings of the Proceedings of the IEEE International Conference on computer vision, 2017, pp. 212–221.
69. Wang, T.; Borji, A.; Zhang, L.; Zhang, P.; Lu, H. A stagewise refinement model for detecting salient objects in images. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017, pp. 4019–4028.
70. Liu, N.; Han, J.; Yang, M.H. Picanet: Learning pixel-wise contextual attention for saliency detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3089–3098.
71. Wang, T.; Zhang, L.; Wang, S.; Lu, H.; Yang, G.; Ruan, X.; Borji, A. Detect globally, refine locally: A novel approach to saliency detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3127–3135.
72. Li, X.; Yang, F.; Cheng, H.; Liu, W.; Shen, D. Contour knowledge transfer for salient object detection. In Proceedings of the Proceedings of the european conference on computer vision (ECCV), 2018, pp. 355–370.
73. Chen, S.; Tan, X.; Wang, B.; Hu, X. Reverse attention for salient object detection. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 234–250.
74. Liu, Y.; Cheng, M.M.; Zhang, X.Y.; Nie, G.Y.; Wang, M. DNA: Deeply supervised nonlinear aggregation for salient object detection. *IEEE Transactions on Cybernetics* **2021**, *52*, 6131–6142.
75. Wu, Z.; Su, L.; Huang, Q. Cascaded partial decoder for fast and accurate salient object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3907–3916.
76. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M. Basnet: Boundary-aware salient object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 7479–7489.
77. Feng, M.; Lu, H.; Ding, E. Attentive feedback network for boundary-aware salient object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 1623–1632.
78. Liu, J.J.; Hou, Q.; Cheng, M.M.; Feng, J.; Jiang, J. A simple pooling-based design for real-time salient object detection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 3917–3926.
79. Zhao, J.X.; Liu, J.J.; Fan, D.P.; Cao, Y.; Yang, J.; Cheng, M.M. EGNet: Edge guidance network for salient object detection. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 8779–8788.
80. Su, J.; Li, J.; Zhang, Y.; Xia, C.; Tian, Y. Selectivity or invariance: Boundary-aware salient object detection. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 3799–3808.

 doi:10.20944/preprints202306.0899.v1

81. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. Icnet for real-time semantic segmentation on high-resolution images. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 405–420.
82. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the Proceedings of the European conference on computer vision (ECCV), 2018, pp. 325–341.
83. Li, H.; Xiong, P.; Fan, H.; Sun, J. Dfanet: Deep feature aggregation for real-time semantic segmentation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9522–9531.