A Causal Debiasing Framework for Unsupervised Salient Object Detection

Xiangru Lin[1,2], Ziyi Wu[1], Guanqi Chen[1], Guanbin Li[1*], Yizhou Yu[2,3]

[1]Sun Yat-sen University [2]The University of Hong Kong [3]Deepwise AI Lab
lin0111@connect.hku.hk, wuzy39@mail2.sysu.edu.cn, chengq26@mail2.sysu.edu.cn,
liguanbin@mail.sysu.edu.cn, yizhouy@acm.org

Abstract

Unsupervised Salient Object Detection (USOD) is a promising yet challenging task that aims to learn a salient object detection model without any ground-truth labels. Self-supervised learning based methods have achieved remarkable success recently and have become the dominant approach in USOD. However, we observed that two distribution biases of salient objects limit further performance improvement of the USOD methods, namely, contrast distribution bias and spatial distribution bias. Concretely, contrast distribution bias is essentially a confounder that makes images with similar high-level semantic contrast and/or low-level visual appearance contrast spuriously dependent, thus forming data-rich contrast clusters and leading the training process biased towards the data-rich contrast clusters in the data. Spatial distribution bias means that the position distribution of all salient objects in a dataset is concentrated on the center of the image plane, which could be harmful to off-center objects prediction. This paper proposes a causal based debiasing framework to disentangle the model from the impact of such biases. Specifically, we use causal intervention to perform deconfounded model training to minimize the contrast distribution bias and propose an image-level weighting strategy that softly weights each image's importance according to the spatial distribution bias map. Extensive experiments on 6 benchmark datasets show that our method significantly outperforms previous unsupervised state-of-the-art methods and even surpasses some of the supervised methods, demonstrating our debiasing framework's effectiveness.

Introduction

Salient Object Detection (SOD) is a fundamental yet challenging computer vision problem that attempts to identify the most visually distinctive parts in an image (Shelhamer, Long, and Darrell 2017; Shetty, Fritz, and Schiele 2018; Simakov et al. 2008; Marchesotti, Cifarelli, and Csurka 2009; Ji et al. 2019; Chen et al. 2020; Li and Yu 2015, 2016). Although the performance of state-of-the-art(SOTA) SOD models has enjoyed a significant improvement in the wave of deep neural networks, training a SOD model usually requires a large amount of pixel-level labeled images, the collection process of which is laborious and time-consuming. USOD is an alternative method to solve this problem, aiming to learn a salient object detection model without using any ground-truth labels. Self-supervised learning based methods have achieved remarkable success recently and have become the dominant approach. The majority of such research typically follow a common pipeline: (1) Traditional handcrafted methods, such as DSR (Li et al. 2013), RBD (Zhu et al. 2014), and MC (Jiang et al. 2013) etc., are utilized to generate pseudo labels that serve as supervisory signals for the training of a more accurate deep neural network based method; (2) Self-supervised training is further performed until the model performance saturates. The performance gain mainly lies in the fact that the pseudo labels contain roughly correct localization and shape of the salient objects. That is, the distribution of visual contrast information and the spatial distribution of salient objects on the image plane are still maintained, guiding the training of a saliency model to capture such dominant information, while being affected as little as possible by the label noise.

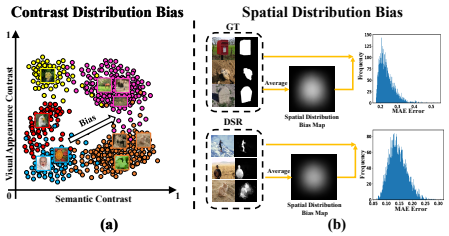

Figure 1: An illustration of the biases discussed in the paper. Figure (a) illustrates the contrast distribution bias where data-rich clusters (the purple and the orange ones) dominate the model training and thus make the model biased towards them; Figure (b) shows the statistical analysis of the spatial distribution bias where the MAE histogram plot displays the mean absolute error between the saliency maps of all images and the spatial distribution bias map.

Although a model trained with the aforementioned pipeline shows competitive performance, we observed that the bias from the distribution of visual contrast information (which we term as contrast distribution bias) forces the model's training to focus on the data-rich visual contrast clusters. Simultaneously, the bias from the spatial distribution of salient objects (which we term as spatial distribution bias) misleads the model training towards predicting the cen-

*Corresponding author is Guanbin Li.

ter area of the image plane as salient. Essentially, visual contrast information, defined as the difference in visual appearance and/or semantics, is the most important factor in visual saliency as demonstrated in the perceptual research (Itti and Koch 2001; Reinagel and Zador 1999). The visual appearance and semantics we mentioned here refer to the low-level features contrast and high-level features contrast respectively. The categorization of visual contrast information has various types (or clusters). However, its distribution in a dataset is not guaranteed to be uniform. In fact, maintaining an absolute uniform distribution of visual contrast information is difficult since controlling the visual contrast distribution in data collection is almost impossible. Thus, this uneven distribution misleads the trained model to make biased predictions towards those data-rich visual contrast clusters. For illustration purposes, we visually separate low and high level visual contrast features as described in later section in two axes and show such contrast distribution bias in MSRA-B dataset in Fig. 1 (a). To better understand the underlying mechanism, we propose a causal graph to explain the causal effect of the contrast distribution bias shown in Fig. 2, where $\boldsymbol{C}$, $\boldsymbol{R}$, $\boldsymbol{I}$, and $\boldsymbol{Y}$ represent the contrast distribution, the saliency specific feature representation, the input image, and the corresponding saliency prediction, respectively. The problem of the causal graph is that the data-rich visual contrast clusters strengthen the supervisory signals for the training of a USOD model through the causal link $\boldsymbol{I} \rightarrow \boldsymbol{R} \rightarrow \boldsymbol{Y}$ whereas the contrast distribution confounds $\boldsymbol{I}$ and $\boldsymbol{Y}$ via the backdoor causal links $\boldsymbol{I} \leftarrow \boldsymbol{C} \rightarrow \boldsymbol{R} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{I} \leftarrow \boldsymbol{C} \rightarrow \boldsymbol{Y}$: the backdoor paths can help to correlate $\boldsymbol{I}$ and $\boldsymbol{Y}$ for some background pixels in $\boldsymbol{I}$ that are not salient at all. Therefore, these causal links mix up the causal effect of $\boldsymbol{I}$ to $\boldsymbol{Y}$, which hinders the learning of the saliency detection model to a certain extent.

On the other hand, salient objects in existing training datasets, such as MSRA-B (Liu et al. 2007) and DUTS (Wang et al. 2017), typically locate at the center of the image plane. In USOD, the objects spatial distribution prior (also called center prior) (Liu et al. 2007; Judd et al. 2009; Goferman, Zelnik-Manor, and Tal 2012) is widely used to improve saliency estimation heuristically. The side effect is that the trained model tends to predict an image's center area to be salient. To illustrate this phenomenon, we accumulate two spatial distribution bias maps from MSRA-B dataset (shown in Fig. 1 (b)) by averaging all predicted saliency maps using DSR and all ground-truth saliency maps respectively. Then, we plot the histogram of the mean absolute error between the accumulated maps and the DSR predicted maps and the ground-truth maps respectively. The smaller the error, the more similar the predicted saliency map ot the ground-truth map is to the spatial distribution bias map. It is obvious that most of the mae errors lean towards the leftmost area (we observed similar phenomenon in DUTS dataset), which indicates that the spatial distribution bias exists and thus, could limit performance improvement.

According to the above analysis, we propose a new causal inference based debiasing framework to disentangle the model from the aforementioned biases. Concretely, we devise a **De-confounded Training** method by backdoor adjustment to remove the contrast distribution bias and propose an **Image-level Weighting Strategy** by calculating the normalized image-level importance that softly weights each image in training to eradicate the spatial distribution bias. Different from the commonly used multi-stage pipeline, we propose a lightweight single-stage method that uses only one traditional handcrafted method. Experiments show that our method is robust to the choice of handcrafted methods.

To sum up, this paper has the following contributions:

- To our best knowledge, we are the first to analyze the contrast distribution and spatial distribution biases in USOD from the causal inference perspective and identify that the contrast distribution bias is a confounder and the spatial distribution bias misleads the model training, which negatively affect the training process.
- We propose a de-confounded training method to remove the confounding effect caused by the contrast distribution bias so that visual contrasts contribute fairly to the final saliency prediction.
- We introduce an image-level weighting strategy that softly weights each image's importance to minimize the misleading impact of the spatial distribution bias.
- Our method outperforms all previous unsupervised methods by a large margin, achieving new state-of-the-art USOD performance across 6 datasets.

Related Works

Unsupervised Salient Object Detection. For deep learning based methods, (Zhang, Han, and Zhang 2017; Zhang et al. 2018; Nguyen et al. 2019; Zhang, Xie, and Barnes 2020) propose to use handcrafted methods as pseudo label producers to train a deep neural network. Specifically, Zhang et al. (Zhang, Han, and Zhang 2017) fused predictions from multiple unsupervised handcrafted methods heuristically using DHSNet. Zhang et al. (Zhang et al. 2018) introduced a noise fitting framework to capture the pseudo label noise among different saliency methods. Nguyen et al. (Nguyen et al. 2019) devised a self-supervised learning based method to refine the pseudo labels from multiple handcrafted methods. However, our work is significantly different from previous works. We propose to eradicate two distribution biases to achieve better pseudo labels in training, which exempts the model learning from such heavy pipeline and motivates us to the single-stage design.

Causal Inference. It has been applied to many computer tasks including natural language processing (Liu et al. 2021; Keith, Jensen, and O'Connor 2020), computer vision (Tang et al. 2020; Zhang et al. 2020a; Wang et al. 2020), Robotics (Ahmed et al. 2021). They mainly apply causal inference to the image/video level tasks while we target at pixel-level task. In our design, we analyze the confounding effect of visual contrast distribution and then, propose a grid confounder set that specifically models the high-level semantic visual contrast and the low-level visual appearance contrast where each grid represents a specific latent type of visual contrast. To our best knowledge, this is the first tailored design for USOD and it establishs a feasible and formal causal framework for interpreting and tackling confounding biases for pixel-level task.

Method

Problem Definition

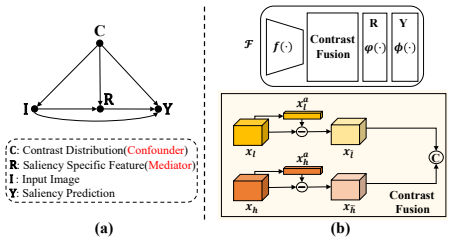

Figure 2: (a) Overview of the proposed causal graph explaining the causal effect of the contrast distribution bias. (b) The proposed strong baseline model $\mathcal{F}$ with explicit visual contrast modeling.

Proposed Strong Baseline Model Settings. Given a set of training images $\mathbb{I} = \{\boldsymbol{I}_i\}_{i=1}^{N_d}$ where N_d is the total number of images, and a backbone network $f(\cdot, \boldsymbol{\theta}_f)$, the extracted feature set is defined as $\mathbb{X} = \{\boldsymbol{X}_i\}_{i=1}^{N_d}$ where $\boldsymbol{X}_i = f(\boldsymbol{I}_i)$. Note that $\boldsymbol{X}_i = \{\boldsymbol{x}_j\}_{j=1}^{N_b}$ where N_b is the number of layers in the backbone network. We first propose a strong baseline model $\mathcal{F}$ shown in Fig. 2(b).

Unlike other USOD methods, we explicitly model the visual contrast representation by designing a contrast fusion block. Concretely, given an input image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, we first compute $\boldsymbol{X} = f(\boldsymbol{I})$. Then, we select low-level visual appearance feature $\boldsymbol{x}_l$ and high-level semantic feature $\boldsymbol{x}_h$ from $\boldsymbol{X}$. The contrast fusion block is defined as follows,

$$\boldsymbol{x}_{\tilde{l}} = \boldsymbol{x}_l - \boldsymbol{x}_l^a; \boldsymbol{x}_{\tilde{h}} = \boldsymbol{x}_h - \boldsymbol{x}_h^a; \boldsymbol{x}_c = [\boldsymbol{x}_{\tilde{h}}, \boldsymbol{x}_{\tilde{l}}] \tag{1}$$

where $\boldsymbol{x}_l^a$ and $\boldsymbol{x}_h^a$ are modeled as the global context features respectively and they are calculated as the mean pooled feature of $\boldsymbol{x}_l$ and $\boldsymbol{x}_h$ respectively. $[\cdot]$ is a concatenation operation. $\boldsymbol{x}_c$ is the final visual contrast enhanced feature. The final saliency prediction map of $\boldsymbol{I}$ is defined as $\boldsymbol{y} = \phi(\varphi(\boldsymbol{x}_c))$ where $\varphi(\cdot, \boldsymbol{\theta}_\varphi)$ learns saliency specific representation R in Fig. 2(a) and $\phi(\cdot, \boldsymbol{\theta}_\phi)$ is the final classifier layer.

Proposed Strong Baseline Model Training. Following (Zhang, Han, and Zhang 2017; Zhang et al. 2018; Nguyen et al. 2019; Zhang, Xie, and Barnes 2020; Zhang et al. 2020b), we first use a traditional handcrafted method such as DSR (Li et al. 2013), MC (Jiang et al. 2013), or RBD (Zhu et al. 2014), to generate pseudo labels for the training of the strong baseline model $\mathcal{F}$. Note that DSR and RBD used center prior, MC without center prior. Then, since the generated pseudo labels are noisy, rather than fixing the pseudo labels for training, we adopt historical moving averages (Nguyen et al. 2019) of predictions to update the pseudo labels after each epoch. Specifically, the moving average function is defined as,

$$\boldsymbol{y}^t = (1-\alpha)CRF(\boldsymbol{y}^{t-1}) + \alpha\boldsymbol{y}^t \tag{2}$$

where $\boldsymbol{y}^t$ is the prediction of $\boldsymbol{I}$ at epoch t and $CRF(\cdot)$ is the conditional random field function (Gupta et al. 2020). α is a blending parameter set to 0.7. The intuition behind this design is to gradually incorporate the fine-grained visual saliency information from $CRF(\cdot)$ while at the same time maintaining stable model predictions. With $\boldsymbol{y}^t$, the pseudo label of $\boldsymbol{I}$ at epoch t is defined as,

$$\boldsymbol{l}^t = \boldsymbol{y}^t \geq \mu^t \tag{3}$$

where $\mu^t = \frac{\gamma}{HW}\boldsymbol{y}_b^{t-1}$ and $\boldsymbol{y}_b^{t-1} = \frac{1}{N_d}\sum_i \boldsymbol{y}_i^{t-1}$. Here, $\boldsymbol{y}_i^{t-1}$ denotes the prediction of the i-th image in $\mathbb{I}$ at epoch $t-1$. μ^t is the moving mean of all prediction scores in all saliency prediction maps in the training set. $\boldsymbol{y}_b^{t-1}$ is the accumulated spatial distribution bias map. γ is a temperature parameter set to 1.5.

Since the pseudo label computed at each epoch t is noisy, we utilize the relaxed F-measure loss $\mathcal{L}_{F_m}^t$ (Nguyen et al. 2019; Zhao et al. 2019b) as the main objective function to supervise the training of the strong baseline model $\mathcal{F}$. This loss function is a linear loss and thus is more robust to noise. The formula is defined as follows,

$$\mathcal{L}_{F_m}^t = 1 - (1+\beta^2)\frac{prec^t * rec^t}{\beta^2 prec^t + rec^t} \tag{4}$$

where $prec^t$ and rec^t are the standard precision and recall between continuous saliency prediction $\boldsymbol{y}^t$ and discrete pseudo label $\boldsymbol{l}^t$. β^2 is set to 0.3 as in the standard evaluation metric.

Besides, to encourage the predicted saliency map having consistent intensities inside the salient region and distinct boundaries at the object edges, following (Qin et al. 2019; Zhang, Xie, and Barnes 2020; Zhang et al. 2020b), we apply a saliency prediction map level IoU loss function $\mathcal{L}_{iou}^t$ and a structure aware edge preserving pixel-level loss function $\mathcal{L}_{edge}^t$ as auxiliary regularizers. Combining them helps the training become robust to noise pseudo labels and forces the model to be aware of salient objects' structure information. The formulas are defined as follows,

$$\mathcal{L}_{iou}^t = 1 - \frac{\sum_{i=1}^H\sum_{j=1}^W \boldsymbol{y}^t(i,j)\boldsymbol{l}^t(i,j)}{\sum_{i=1}^H\sum_{j=1}^W \left[\boldsymbol{y}^t(i,j) + \boldsymbol{l}^t(i,j) - \boldsymbol{y}^t(i,j)\boldsymbol{l}^t(i,j)\right]} \tag{5}$$

$$\mathcal{L}_{edge}^t = \sum_{i,j}\sum_{d\in\overrightarrow{x},\overrightarrow{y}} \Psi(\left|\partial_d \boldsymbol{y}^t(i,j)\right| e^{-\kappa|\partial_d \boldsymbol{I}(i,j)|}) \tag{6}$$

where $\Psi(\cdot)$ is the Charbonnier penalty formula $\Psi(s) = \sqrt{s^2 + 1e-6}$. (i,j) represents a pixel coordinate, and d indexes over the partial derivatives (first order derivatives) in $\overrightarrow{x}$ and $\overrightarrow{y}$ directions in the image plane. κ is set to 10 as in (Qin et al. 2019; Zhang, Xie, and Barnes 2020; Zhang et al. 2020b).

Finally, the overall objective function for the proposed strong baseline model at epoch t is summarized as follows,

$$\mathcal{L}^t = w_1\mathcal{L}_{F_m}^t + w_2\mathcal{L}_{iou}^t + w_3\mathcal{L}_{edge}^t \tag{7}$$

where $w_1, w_2, w_3 \in \mathbb{R}$ are loss weights.

Contrast Distribution Bias

Analysis. Although the strong baseline model $\mathcal{F}$ trained with the previous pipeline exhibits favorable performance, the performance gain primarily comes from the dominant visual contrast clusters hidden in the training set, as shown in Fig. 1 (a). This phenomenon is, in fact, a two-edged

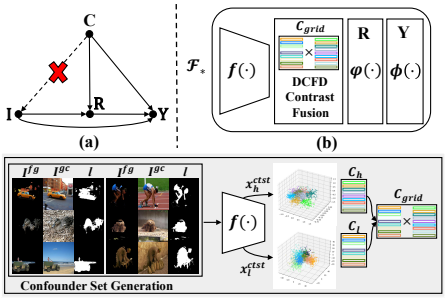

Figure 3: Overview of the proposed de-confounded causal graph with causal intervention. DCFD Contrast Fusion represents the de-confounded process defined in equation(8).

sword: on the one hand, it guides the model training to capture the most discriminative visual contrast representation in the noisy pseudo labeled dataset so that the learned model could reach a better local minimum; on the other hand, for those visual contrast clusters with less data, their contribution to the model training will be down-weighted or even ignored, thus making the model biased and limiting the model's performance. This can be interpreted from a causal inference perspective. We propose a causal graph using Pearl's graphical model (Pearl 2009) in Fig. 2 to explain such phenomenon.

Concretely, (1) the causal link $\boldsymbol{C} \rightarrow \boldsymbol{I}$: the contrast distribution $\boldsymbol{C}$ is determined by the data collection process and once collected, for each image $\boldsymbol{I}$, $\boldsymbol{C}$ determines the content of the image; (2) the causal link $\boldsymbol{C} \rightarrow \boldsymbol{Y}$: the contrast distribution affects the saliency prediction of an image via the corresponding pseudo label generated by a traditional method, such as DSR, MC, or RBD; (3) the causal link $\boldsymbol{I} \rightarrow \boldsymbol{R} \rightarrow \boldsymbol{Y}$: $\boldsymbol{R}$ is a saliency specific representation of $\boldsymbol{I}$ obtained by a USOD model, which serves a mediator before the final classification layer of the USOD model; (4) the causal link $\boldsymbol{C} \rightarrow \boldsymbol{R} \rightarrow \boldsymbol{Y}$: the contrast distribution generates the pseudo label to supervise the training of the saliency specific representation, which resembles the pseudo label training process whereas (2) resembles the initial pseudo label generation process; (5) $\boldsymbol{I} \rightarrow \boldsymbol{Y}$ is the direct causal effect that we aim to achieve. According to (Pearl, Glymour, and Jewell 2016), contrast distribution $\boldsymbol{C}$ is a confounder which is the common cause of the input image $\boldsymbol{I}$ and its corresponding saliency prediction $\boldsymbol{Y}$. The positive effect of those data-rich visual contrast clusters follows the causal link $\boldsymbol{I} \rightarrow \boldsymbol{R} \rightarrow \boldsymbol{Y}$ to learn a contrast aware discriminative USOD model while the negative effect of those clusters forces some non-salient background pixels in $\boldsymbol{I}$ to be salient $\boldsymbol{Y}$ following the backdoor causal links ($\boldsymbol{I} \leftarrow \boldsymbol{C} \rightarrow \boldsymbol{R} \rightarrow \boldsymbol{Y}$ and $\boldsymbol{I} \leftarrow \boldsymbol{C} \rightarrow \boldsymbol{Y}$).

Intervened Causal Graph. To address the aforementioned confounding effect, we propose an improved causal graph shown in Fig. 3. Specifically, to remove the dominant contribution made by those data-rich visual contrast clusters, an intuitive idea is to make each cluster contribute to the saliency prediction fairly, that is changing the $P(Y|I)$ to $P(Y|do(\boldsymbol{I})) = \sum_{\boldsymbol{c}} P(\boldsymbol{Y}|\boldsymbol{I}, \boldsymbol{R}, \boldsymbol{c}) P(\boldsymbol{c})$. This is termed as backdoor adjustment (Pearl, Glymour, and Jewell 2016). Since we study a single cause confounder $\boldsymbol{C}$, this intervention is performed by (1) cutting off the $\boldsymbol{C} \rightarrow \boldsymbol{I}$ link and (2) then stratifying $\boldsymbol{C}$ into pieces to construct a confounder set which contains the distribution information of the visual contrast representation. To this end, we implement the stratification of $\boldsymbol{C}$ as $\boldsymbol{C} = \{\boldsymbol{c}_i\}_{i=1}^{N_c}$ where $\boldsymbol{c}_i \in \mathbb{R}^{D_c}$ and N_c is a hyperparameter representing the size of the confounder set. Since the number of visual contrast clusters is large in real world and there are no ground-truth visual contrast cluster information in the training set, we use K-Means++ with Principle Component Analysis (PCA) to learn the confounder set $\boldsymbol{C}$ so that each $\boldsymbol{c}_i$ represents a form of visual contrast cluster (type). Each cluster $\boldsymbol{c}_i$ is set to the mean feature of each cluster in K-Means++.

Then, how do we choose the feature representation for $\boldsymbol{C}$? As mentioned previously, we model the visual contrast of an image $\boldsymbol{I}$ as the difference between the salient object region and the global context. Concretely, (1) we forward an image $\boldsymbol{I}$ to the backbone network $f(\cdot)$ to obtain the feature set $\boldsymbol{X} = \{\boldsymbol{x}_j\}_{j=1}^{N_b}$; (2) we then perform mean pooling on $\boldsymbol{x}_j$ and treat it as the global context feature $\boldsymbol{x}_j^{gc}$ for layer j and apply the generated pseudo label (denoted as $\boldsymbol{l}$ in Fig. 3) to $\boldsymbol{x}_j$ to extract the feature of the salient object region at layer j, which is also mean pooled and is denoted as $\boldsymbol{x}_j^{fg}$; (3) finally, the visual contrast representation is defined as $\boldsymbol{x}_j^{cst} = \boldsymbol{x}_j^{fg} - \boldsymbol{x}_j^{gc}$. Since visual contrast appears in the form of both low-level visual appearance feature and the high-level semantic feature, we model low-level visual contrast as $\boldsymbol{x}_l^{cst} = \boldsymbol{x}_l^{fg} - \boldsymbol{x}_l^{gc}$ and high-level visual contrast as $\boldsymbol{x}_h^{cst} = \boldsymbol{x}_h^{fg} - \boldsymbol{x}_h^{gc}$, where h is selected from the higher layer in $f(\cdot)$ and l is selected as the lower layer in $f(\cdot)$ respectively.

To this end, we learn confounder sets $\boldsymbol{C}_h$ and $\boldsymbol{C}_l$ for high-level visual semantic contrast and low-level visual appearance contrast respectively. Considering the fact that a region in an image can be visually salient in the form of low-level visual contrast, high-level semantic contrast or both, we concatenate the features in $\boldsymbol{C}_h$ and $\boldsymbol{C}_l$ to form a grid confounder set $\boldsymbol{C}_{grid} = \{\boldsymbol{c}_{i,j}\}_{i=1}^{N_c^h \times N_c^l}$ where $\boldsymbol{c}_{i,j} = [\boldsymbol{c}_{h,i}, \boldsymbol{c}_{l,j}] \in \mathbb{R}^{D_c^h + D_c^l}$. Here, $\boldsymbol{c}_{h,i}$ represents the i-th cluster center in $\boldsymbol{c}_h$ and $\boldsymbol{c}_{l,j}$ represents the j-th cluster center in $\boldsymbol{c}_l$. N_c^h and N_c^l denote the size of the high-level and low-level visual contrast confounder set respectively. In the implementation, calculating $P(Y|do(\boldsymbol{I}))$ requires multiple forward passes of all $\boldsymbol{c}$. To reduce the computational cost, we approximate the summation at the feature level as in (Vaswani et al. 2017; Zhang et al. 2020a),

$$P(\boldsymbol{Y}|do(\boldsymbol{I})) \approx P\left(\boldsymbol{Y}|\boldsymbol{I}, \boldsymbol{R} = \sum_{\boldsymbol{c}} g(\boldsymbol{x} = f(\boldsymbol{I}), \boldsymbol{c}) P(\boldsymbol{c})\right) \tag{8}$$

where $\boldsymbol{c}$ represents (i,j)-th cluster $\boldsymbol{c}_{i,j}$ in $\boldsymbol{C}_{grid}$. The approximation is achieved by the Normalized Weighted Geometric Mean (Xu et al. 2015). $P(\boldsymbol{c}) = \frac{|\boldsymbol{c}|}{\sum_{i,j}|\boldsymbol{c}_{i,j}|}$ and $|\boldsymbol{c}|$ is the number of samples in cluster $\boldsymbol{c}$. In this way, the contrast distribution $\boldsymbol{C}$ in Fig. 3 (a) is no longer correlated with $\boldsymbol{I}$ and thus, this causal intervention makes $\boldsymbol{I}$ contribute fairly to

$\boldsymbol{Y}$'s prediction by incorporating every contrast cluster $\boldsymbol{c}_{i,j}$. We implement $g(\cdot)$ as a soft attention over $\boldsymbol{C}_{grid}$,

$$\begin{aligned} g(\boldsymbol{x}, \boldsymbol{C}_{grid}) &= \sum_{\boldsymbol{c}} g(\boldsymbol{x}, \boldsymbol{c}) = Attention(\boldsymbol{x}, \boldsymbol{C}_{grid}) \\ &= softmax(\frac{(\boldsymbol{W}_1\boldsymbol{x})^T(\boldsymbol{W}_2\boldsymbol{C}_{grid})}{\sqrt{D_h}})\boldsymbol{C}_{grid} \end{aligned} \tag{9}$$

where $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ are learnable parameters to project $\boldsymbol{x}$ and $\boldsymbol{C}_{grid}$ into a joint space. $\sqrt{D_h}$ is a constant scaling factor for feature normalization. Since $g(\cdot)$ is essentially a soft-attention between feature tensor $\boldsymbol{x}$ and confounder set $\boldsymbol{C}_{grid}$, we implement $g(\cdot)$ as a multi-head attention layer as in (Vaswani et al. 2017). Compared to the strong baseline model $\mathcal{F}$, the final saliency prediction $\boldsymbol{y}$ of the image $\boldsymbol{I}$ is achieved by the following formula,

$$\boldsymbol{y} = \phi(\varphi([\boldsymbol{x}, \sum_{\boldsymbol{c}} g(\boldsymbol{x}, \boldsymbol{c})])) \tag{10}$$

where $[\cdot]$ is the concatenation operator.

Spatial Distribution Bias

Besides the aforementioned Contrast Distribution Bias, we observed that the spatial objects' distribution generally follows a Gaussian fall-off map centered at the center of the image plane, which is reasonable since the training dataset contains a large amount of near-center salient objects and moreover, traditional USOD methods typically use center-prior to generate pseudo labels. While such a bias can improve saliency results for many images, especially when the supervision signal is weak in USOD, they can fail when a salient object is off-center. To mitigate such bias, we propose an image-level weighting strategy that softly weights each image's importance according to the mean absolute error between the saliency prediction of an image and the calculated spatial distribution bias map. Concretely, given an image $\boldsymbol{I}$ and its corresponding saliency prediction $\boldsymbol{y}^t$ at epoch t, the image weight η^t for $\boldsymbol{I}$ at epoch t is defined as follows,

$$\eta^t = \frac{exp(MeanPool(\boldsymbol{y}^t - \boldsymbol{y}_b^t))^{\frac{1}{T'}}}{\sum_j exp(MeanPool((\boldsymbol{y}_j^t - \boldsymbol{y}_b^t)))^{\frac{1}{T'}}} \tag{11}$$

where $\boldsymbol{y}_b^t = \frac{1}{N_d}\sum_i \boldsymbol{y}_i^t$. $MeanPool(\cdot)$ calculates the mean value of the input. $T^{'}$ is a temperature parameter set to 1.5. Enlarging $T^{'}$ will result in higher mean value of all image weights. Note that our image weighting strategy is different from focal loss in that focal loss balances the confidence of each training sample while our method balances samples according to the distribution of each sample with respect to the spatial distribution bias map. Although both methods reweight samples, our method explicitly alleviates the influence of the spatial distribution bias in training.

Therefore, the updated overall objective function $\mathcal{L}_*^t$ for epoch t is defined as follows,

$$\mathcal{L}_*^t = \sum\nolimits_i \eta_i^t \mathcal{L}_{F_m,i}^t + w_2\mathcal{L}_{iou}^t + w_3\mathcal{L}_{edge}^t \tag{12}$$

where η_i^t denotes the calculated image weight for i-th image. Finally, we define the de-confounded model trained with $\mathcal{L}_*^t$ as $\mathcal{F}_*$.

Experiments

Experiments Setup

Settings. The aim of our proposed causal debiasing framework is 1) to unravel the ever-underestimated biases in USOD and 2) to prove that such biases exist and mitigating such biases could boost the prediction performance. Thus, we focus on improving the performance of USOD model given pseudo labels generated by a handcrafted method. Following (Zhang, Han, and Zhang 2017; Nguyen et al. 2019; Zhang, Xie, and Barnes 2020), we test our framework on DSR (Li et al. 2013), RBD (Zhu et al. 2014), and MC (Jiang et al. 2013) respectively and show that our framework could improve each one of them by a large margin. Additionally, to make fair comparisons with previous state-of-the-art method (Nguyen et al. 2019) trained on MSRA-B (Liu et al. 2007) dataset, following (Nguyen et al. 2019), we adopt DRN-network (Yu, Koltun, and Funkhouser 2017) as $f(\cdot)$; to make fair comparisons with previous state-of-the-art method (Zhang, Xie, and Barnes 2020) trained on DUTS (Wang et al. 2017) dataset, we utilize ResNet-50 (He et al. 2016) based DeepLabV2 (Chen et al. 2018) as $f(\cdot)$.

Evaluation Metrics & Implementation Details. Following (Zhang, Han, and Zhang 2017; Nguyen et al. 2019; Zhang, Xie, and Barnes 2020), we report three evaluation metrics, namely Mean Absolute Error ($\mathcal{M}$), mean F-measure (F_β), E-measure (E_ξ) (Fan et al. 2018). For more implementation details, please refer to **the supplementary document**.

Comparison to state-of-the-art methods

Quantitative Comparisons. According to Tab. 1, we construct two groups of experiments to verify the effectiveness of our proposed causal debiasing framework. The first group of experiments is termed as $Group_M$ consisting of experiments $24-29$ trained on the MSRA-B training set. The second group termed as $Group_D$ is composed of experiments $30-35$ trained on the DUTS training set. The three handcrafted methods are tested individually in each group. Also, the strong baseline model (termed as Strong Baseline) is trained using $\mathcal{L}^t$ while the full model (termed as Ours DCFD) is trained with $\mathcal{L}_*^t$.

For models trained on the MSRA-B dataset, the previous best performing USOD method is the DeepUSPS-Fuse (Nguyen et al. 2019)(experiment 19). When both DeepUSPS and our method train on single handcrafted method on MSRA-B dataset using RBD, MC, and DSR, respectively, our method outperforms DeepUSPS by a large margin (Averagely, surpass DeepUSPS-RBD by $6.4\%(E_\xi)$, $11.6\%(F_\beta)$, $4.1\%(\mathcal{M})$; surpass DeepUSPS-MC by $3.8\%(E_\xi)$, $8.0\%(F_\beta)$, $2.8\%(\mathcal{M})$; surpass DeepUSPS-DSR by $6.4\%(E_\xi)$, $11.1\%(F_\beta)$, $4.3\%(\mathcal{M})$). When DeepUSPS is trained for all four stages fused with four handcrafted methods, our method performs comparably well against DeepUSPS-Fuse (for RBD, $-0.5\%(E_\xi)$, $-0.1\%(F_\beta)$, $+0.05\%(\mathcal{M})$; For MC, $-0.4\%(E_\xi)$, $+0.01\%(F_\beta)$, $+0.17\%(\mathcal{M})$; For DSR, $-0.2\%(E_\xi)$, $+0.86\%(F_\beta)$, $+0.18\%(\mathcal{M})$). Moreover, our method could easily adapt to training on larger datasets

Table 1: Quantitative comparison with state-of-the-art SOD methods on 6 datasets in terms of E-measure $E_\xi \uparrow$, mean F-measure $F_\beta \uparrow$, and MAE $\mathcal{M} \downarrow$. $\uparrow$ and $\downarrow$ indicate larger and smaller is better, respectively. The best performing of fully-supervised and Weakly-/Un-supervised method is marked in **bold**, respectively.

Supervision	Method	Training Set	ID	DUTS-TE			ECSSD			DUT-OMRON			HKU-IS			PASCAL-S			MSRA-B		
				$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$
Fully Supervised	DGRL (Wang et al. 2018)	DUTS	1	0.893	0.794	0.049	0.946	0.906	0.041	0.856	0.733	0.062	0.949	0.890	0.036	0.888	0.819	0.072	0.936	0.885	0.043
	PiCANet (Liu, Han, and Yang 2018)	DUTS	2	0.866	0.749	0.053	0.926	0.885	0.046	0.842	0.710	0.068	0.938	0.870	0.042	0.858	0.789	0.078	-	-	-
	F3Net (Wei, Wang, and Huang 2020)	DUTS	3	**0.913**	**0.840**	**0.035**	**0.946**	**0.925**	**0.033**	**0.876**	**0.766**	**0.053**	**0.958**	**0.910**	**0.028**	**0.894**	**0.835**	**0.061**	-	-	-
	NLDF (Luo et al. 2017)	DUTS	4	-	-	-	0.912	0.878	0.063	0.817	0.684	0.080	0.929	0.874	0.048	0.840	0.769	0.098	0.934	0.889	0.048
	PoolNet (Liu et al. 2019)	DUTS	5	0.894	0.799	0.041	0.943	0.910	0.042	0.863	0.739	0.056	0.952	0.894	0.033	0.873	0.806	0.072	-	-	-
	BASNet (Qin et al. 2019)	DUTS	6	0.879	0.791	0.047	0.921	0.880	0.037	0.869	0.756	0.056	0.946	0.895	0.032	0.852	0.771	0.076	-	-	-
	AFNet (Feng, Lu, and Ding 2019)	DUTS	7	0.893	0.793	0.045	0.941	0.908	0.042	0.859	0.739	0.057	0.947	0.888	0.036	0.884	0.820	0.070	-	-	-
	MSNet (Wu et al. 2019)	DUTS	8	0.893	0.799	0.046	0.948	0.914	0.038	0.864	0.742	0.056	0.949	0.892	0.034	0.891	0.831	0.068	-	-	-
	SCRN (Wu, Su, and Huang 2019)	DUTS	9	0.899	0.809	0.039	0.942	0.918	0.037	0.869	0.746	0.056	0.953	0.896	0.034	0.887	0.827	0.063	-	-	-
	EGNet (Zhao et al. 2019a)	DUTS	10	0.904	0.815	0.039	0.947	0.920	0.037	0.874	0.756	0.053	0.956	0.901	0.031	0.876	0.817	0.074	-	-	-
	MINet (Pang et al. 2020)	DUTS	11	0.913	0.828	0.037	0.953	0.924	0.033	0.873	0.755	0.056	0.960	0.909	0.029	0.897	0.829	0.064	-	-	-
Weakly-/Un-supervised	RBD (Zhu et al. 2014)	-	12	0.709	0.501	0.156	0.787	0.676	0.171	0.719	0.525	0.146	0.811	0.676	0.143	0.716	0.591	0.199	0.873	0.787	0.115
	DSR (Li et al. 2013)	-	13	0.716	0.511	0.148	0.787	0.690	0.171	0.721	0.524	0.139	0.807	0.674	0.143	0.705	0.581	0.204	0.862	0.773	0.121
	MC (Jiang et al. 2013)	-	14	0.722	0.521	0.198	0.787	0.699	0.202	0.728	0.533	0.186	0.804	0.677	0.185	0.711	0.596	0.230	0.871	0.799	0.148
	SBF (Zhang, Han, and Zhang 2017)	MSRA10K	15	-	-	-	0.875	0.809	0.090	0.770	0.609	0.110	-	-	-	0.790	0.695	0.133	-	-	-
	WSI (Li, Xie, and Lin 2018)	MSRA-B+HKUIS	16	0.770	0.614	0.115	0.853	0.798	0.110	0.776	0.622	0.101	0.878	0.806	0.086	0.771	0.693	0.149	0.919	0.871	0.067
	WSS (Wang et al. 2017)	DUTS	17	0.790	0.645	0.101	0.867	0.820	0.105	0.765	0.600	0.110	0.892	0.812	0.081	0.789	0.712	0.140	0.906	0.846	0.077
	MNL (Zhang et al. 2018)	MSRA-B	18	-	-	-	0.906	0.874	0.069	0.821	0.683	0.076	0.932	0.874	0.047	0.846	0.792	0.091	0.932	0.881	0.053
	DeepUSPS-Fuse (Nguyen et al. 2019)	MSRA-B	19	0.840	0.730	0.072	0.903	0.875	0.064	0.839	0.715	0.069	0.933	0.880	0.043	0.828	0.770	0.107	0.938	0.896	0.042
	DeepUSPS-RBD (Nguyen et al. 2019)	MSRA-B	20	0.770	0.612	0.103	0.842	0.782	0.102	0.751	0.564	0.116	0.862	0.765	0.083	0.751	0.641	0.153	0.873	0.802	0.084
	DeepUSPS-MC (Nguyen et al. 2019)	MSRA-B	21	0.799	0.645	0.092	0.869	0.810	0.089	0.781	0.609	0.098	0.893	0.800	0.069	0.788	0.696	0.135	0.896	0.828	0.071
	DeepUSPS-DSR (Nguyen et al. 2019)	MSRA-B	22	0.775	0.623	0.105	0.841	0.791	0.105	0.766	0.590	0.112	0.864	0.776	0.085	0.763	0.663	0.149	0.874	0.807	0.085
	EDNS (Zhang, Xie, and Barnes 2020)	DUTS	23	**0.847**	**0.735**	**0.065**	**0.906**	**0.872**	**0.068**	**0.821**	**0.682**	**0.076**	**0.933**	**0.874**	**0.046**	**0.845**	**0.790**	**0.094**	**0.932**	**0.880**	**0.051**
	Strong Baseline(RBD)	MSRA-B	24	0.749	0.687	0.079	0.827	0.829	0.091	0.780	0.693	0.072	0.861	0.849	0.063	0.766	0.733	0.126	0.903	0.884	0.054
	Strong Baseline(DSR)	MSRA-B	25	0.788	0.719	0.074	0.860	0.854	0.081	0.805	0.714	0.068	0.893	0.871	0.054	0.789	0.744	0.119	0.921	0.896	0.047
	Strong Baseline(MC)	MSRA-B	26	0.746	0.688	0.079	0.829	0.833	0.091	0.770	0.689	0.073	0.858	0.848	0.064	0.765	0.732	0.126	0.903	0.886	0.054
	Ours DCFD(RBD)	MSRA-B	27	0.826	0.732	0.069	0.892	0.867	0.066	0.833	0.718	0.067	0.922	0.880	0.045	0.829	0.768	0.104	0.932	0.895	0.043
	Ours DCFD(DSR)	MSRA-B	28	**0.832**	**0.744**	**0.068**	**0.900**	**0.880**	**0.064**	**0.838**	**0.731**	**0.064**	**0.926**	**0.887**	**0.044**	**0.830**	**0.773**	**0.105**	**0.938**	**0.903**	**0.041**
	Ours DCFD(MC)	MSRA-B	29	0.828	0.729	0.069	0.900	0.873	0.063	0.831	0.718	0.067	0.926	0.881	0.044	0.833	0.772	0.102	0.936	0.897	0.042
	Strong Baseline(RBD)	DUTS	30	0.832	0.737	0.072	0.897	0.877	0.070	0.827	0.687	0.075	0.923	0.882	0.048	0.844	0.784	0.100	0.925	0.887	0.050
	Strong Baseline(DSR)	DUTS	31	0.843	0.747	0.070	0.897	0.877	0.072	0.834	0.697	0.074	0.927	0.884	0.048	0.841	0.783	0.102	0.925	0.887	0.050
	Strong Baseline(MC)	DUTS	32	0.833	0.750	0.071	0.887	0.875	0.075	0.826	0.696	0.075	0.917	0.884	0.050	0.835	0.791	0.104	0.917	0.886	0.053
	Ours DCFD(RBD)	DUTS	33	0.837	0.755	0.066	0.905	0.882	0.064	0.832	0.705	0.071	0.927	0.886	0.045	0.852	0.794	0.092	0.929	0.888	0.047
	Ours DCFD(DSR)	DUTS	34	**0.855**	**0.764**	**0.064**	**0.915**	**0.888**	**0.059**	**0.837**	**0.710**	**0.070**	**0.935**	**0.889**	**0.042**	**0.860**	**0.795**	**0.090**	**0.930**	**0.888**	**0.045**
	Ours DCFD(MC)	DUTS	35	0.846	0.759	0.067	0.905	0.883	0.065	0.827	0.700	0.075	0.928	0.887	0.045	0.852	0.793	0.096	0.924	0.889	0.049

such as DUTS, and the performance is clearly superior to DeepUSPS-Fuse (For instance, DSR, $+0.86\%(E_\xi)$, $+1.13\%(F_\beta)$, $+0.50\%(\mathcal{M})$). The advantages of our method over DeepUSPS is that DeepUSPS requires training with four handcrafted methods and four stages while our method trains on a single handcrafted method with only one stage. It is much simpler and computationally efficient.

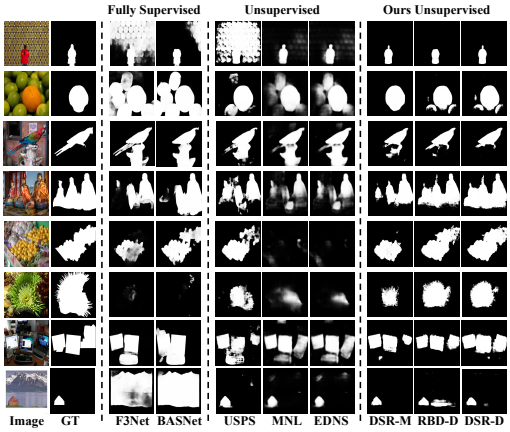

Figure 4: Qualitative Examples of the comparisons between our method and other methods. DSR-M denotes the model trained on MSRA-B dataset using DSR as the handcrafted method; RBD-D represents the model trained on DUTS dataset using RBD as the handcrafted method; DSR-D is the model trained on DUTS dataset using DSR.

For models trained on the DUTS dataset, the previous state-of-the-art USOD method is EDNS (Zhang, Xie, and Barnes 2020)(experiment 23) trained using multiple handcrafted methods together. According to experiments in $Group_D$ and experiment 23 in Tab. 1, it is clear that the performances of our strong baseline model $\mathcal{F}$ using different handcrafted methods are already competitive to that of the previous state-of-the-art USOD methods. Further, by comparing our full model with EDNS in experiments $23, 33, 34, 35$, our proposed models using different handcrafted methods outperform previous SOTA by a clear margin. Specifically, our full model using DSR achieves new SOTA USOD performance in terms of all evaluation metrics, averagely increasing the previous SOTA by $1.0\%, 1.3\%$ in terms of E_ξ and F_β and decreasing $\mathcal{M}$ by 0.5%. Moreover, our best performing DSR method in experiment 34 is even competitive to the fully supervised methods such as PiCANet(experiment 2) and NLDF(experiment 4), demonstrating the superiority of our proposed causal framework.

Qualitative Comparisons. Since there are no ground-truth annotations for the visual contrast clusters and the spatial distribution bias, we select representative examples to qualitatively verify the performance of our method as shown in Fig. 4. As is clear in the examples, the two biases typically co-exist in images. By overfitting to the high-level semantic contrast and ignoring off-center pixels, existing fully supervised methods fail to detect the person located at the bottom of the image in row 1, the orange located at the bottom left of the image in row 2, and the monitor located at the rightmost area of the image in row 7. By overfitting to the center area in training, most methods fail to detect the leftmost statue in row 4. Similar phenomenon can be observed in row 3 where the pillar is not supposed to be detected. These representatives prove that the existence of the two biases and spatial distribution bias exist and our causal debiasing framework could alleviate such biases.

Ablation Study

Component Effectiveness. According to Tab. 2, by comparing experiments $1, 2, 5, 6, 9, 10$ on the MSRA-B dataset and experiments $13, 14, 17, 18, 21, 22$ on the DUTS dataset, the proposed Contrast Fusion Block improves the vanilla baseline by a clear margin, improving by $2.96\%(E_\xi)$, $3.2\%(F_\beta)$, $0.43\%(\mathcal{M})$ for MSRA-B and $5.44\%(E_\xi)$, $2.66\%(F_\beta)$, $0.98\%(\mathcal{M})$ for DUTS, indicating that our baseline model $\mathcal{F}$ is a strong baseline. By replacing CFB with the DCFD CFB in experiments $3, 7, 11$ on MSRA-B dataset and experiments $15, 19, 23$ on the DUTS dataset, the performances are further improved, improving by $2.03\%(E_\xi)$,

Table 2: Ablation Study of our proposed causal debiasing framework on 6 datasets. CFB denotes Contrast Fusion Block; DCFD CFB denotes De-confounded Contrast Fusion Block; IWS represents Image Weighting Strategy. The best method is in **bold**.

Method			Training Set	Handcrafted Method	ID	DUTS-TE			ECSSD			DUT-OMRON			HKU-IS			PASCAL-S			MSRA-B		
CFB	DCFD CFB	IWS				$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$
			MSRA-B	RBD	1	0.697	0.638	0.083	0.796	0.795	0.097	0.735	0.648	0.078	0.819	0.807	0.073	0.749	0.704	0.125	0.870	0.850	0.066
√					2	0.749	0.687	0.079	0.827	0.829	0.091	0.780	0.693	0.072	0.861	0.849	0.063	0.766	0.733	0.126	0.903	0.884	0.054
	√				3	0.781	0.710	0.075	0.853	0.849	0.083	0.805	0.709	0.070	0.885	0.865	0.057	0.791	0.750	0.118	0.917	0.895	0.050
	√	√			4	0.826	0.732	0.069	0.892	0.867	0.066	0.833	0.718	0.067	0.922	0.880	0.045	0.829	0.768	0.104	0.932	0.895	0.043
				DSR	5	0.745	0.687	0.077	0.844	0.838	0.081	0.775	0.688	0.070	0.863	0.846	0.061	0.770	0.724	0.120	0.904	0.880	0.054
√					6	0.788	0.719	0.074	0.860	0.854	0.081	0.805	0.714	0.068	0.893	0.871	0.054	0.789	0.744	0.119	0.921	0.896	0.047
	√				7	0.811	0.729	0.072	0.881	0.868	0.074	0.826	0.723	0.066	0.909	0.879	0.050	0.810	0.759	0.113	0.929	0.899	0.045
	√	√			8	**0.832**	**0.744**	**0.068**	**0.900**	**0.880**	**0.064**	**0.838**	**0.731**	**0.064**	**0.926**	**0.887**	**0.044**	**0.830**	**0.773**	**0.105**	**0.938**	**0.903**	**0.041**
				MC	9	0.703	0.641	0.083	0.810	0.803	0.091	0.734	0.647	0.078	0.825	0.811	0.071	0.768	0.714	0.118	0.873	0.852	0.066
√					10	0.746	0.688	0.079	0.829	0.833	0.091	0.770	0.689	0.073	0.858	0.848	0.064	0.765	0.732	0.126	0.903	0.886	0.054
	√				11	0.768	0.704	0.076	0.849	0.848	0.085	0.785	0.698	0.072	0.876	0.860	0.060	0.786	0.749	0.119	0.912	0.892	0.052
	√	√			12	0.828	0.729	0.069	0.900	0.873	0.063	0.831	0.718	0.067	0.926	0.881	0.044	0.833	0.772	0.102	0.936	0.897	0.042
			DUTS	RBD	13	0.742	0.699	0.079	0.825	0.836	0.092	0.769	0.672	0.076	0.854	0.847	0.066	0.758	0.731	0.121	0.889	0.865	0.062
√					14	0.832	0.737	0.072	0.897	0.877	0.070	0.827	0.687	0.075	0.923	0.882	0.048	0.844	0.784	0.100	0.925	0.887	0.050
	√				15	0.832	0.738	0.071	0.898	0.880	0.069	0.831	0.694	0.073	0.922	0.883	0.048	0.843	0.786	0.099	0.925	0.887	0.049
	√	√			16	0.837	0.755	0.066	0.905	0.882	0.064	0.832	0.705	0.071	0.927	0.886	0.045	0.852	0.794	0.092	0.929	0.888	0.047
				DSR	17	0.787	0.734	0.073	0.858	0.860	0.082	0.793	0.690	0.073	0.885	0.868	0.058	0.790	0.756	0.111	0.903	0.874	0.057
√					18	0.843	0.747	0.070	0.897	0.877	0.072	0.834	0.697	0.074	0.927	0.884	0.048	0.841	0.783	0.102	0.925	0.887	0.050
	√				19	0.844	0.750	0.069	0.900	0.880	0.070	0.837	0.703	0.072	0.928	0.887	0.047	0.844	0.786	0.100	0.925	0.888	0.049
	√	√			20	**0.855**	**0.764**	**0.064**	**0.915**	**0.888**	**0.059**	**0.837**	**0.710**	**0.070**	**0.935**	**0.889**	**0.042**	**0.860**	**0.795**	**0.090**	**0.930**	**0.888**	**0.045**
				MC	21	0.764	0.714	0.077	0.842	0.849	0.088	0.764	0.666	0.078	0.866	0.856	0.063	0.770	0.748	0.118	0.892	0.867	0.061
√					22	0.833	0.750	0.071	0.887	0.875	0.075	0.826	0.696	0.075	0.917	0.884	0.050	0.835	0.791	0.104	0.917	0.886	0.053
	√				23	0.844	0.750	0.069	0.902	0.882	0.067	0.832	0.697	0.074	0.928	0.886	0.046	0.845	0.787	0.099	0.925	0.890	0.049
	√	√			24	0.846	0.759	0.067	0.905	0.883	0.065	0.827	0.700	0.075	0.928	0.887	0.045	0.852	0.793	0.096	0.924	0.889	0.049

Table 3: Ablation Study of Grid Confounder Set Contrast Feature Selection on 6 datasets. The best method is in **bold**.

Handcrafted Method	Training Set	Backbone Network	ID	Low-level Layer C_l	High-level Layer C_h	DUTS-TE			ECSSD			DUT-OMRON			HKU-IS			PASCAL-S			MSRA-B		
						$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$	$E_\xi\uparrow$	$F_\beta\uparrow$	$\mathcal{M}\downarrow$
DSR	DUTS	ResNet-50	1	res2	res5	**0.855**	**0.764**	**0.064**	**0.915**	**0.888**	**0.059**	**0.837**	**0.710**	**0.070**	**0.935**	**0.889**	**0.042**	**0.860**	**0.795**	**0.090**	**0.930**	**0.888**	**0.045**
			2	res3	res5	0.856	0.768	0.063	0.914	0.889	0.059	0.838	0.712	0.068	0.932	0.889	0.043	0.859	0.797	0.090	0.931	0.888	0.045
			3	res2	res4	0.840	0.763	0.065	0.897	0.880	0.066	0.832	0.719	0.067	0.926	0.889	0.044	0.845	0.786	0.096	0.927	0.892	0.045
			4	res3	res4	0.821	0.760	0.065	0.882	0.874	0.070	0.817	0.716	0.065	0.908	0.882	0.048	0.821	0.776	0.102	0.923	0.891	0.046
	MSRA-B	DRN	5	layer2	layer8	0.829	0.743	0.068	0.897	0.876	0.066	0.837	0.730	0.064	0.923	0.885	0.045	0.821	0.764	0.107	0.936	0.902	0.041
			6	layer3	layer8	**0.832**	**0.744**	**0.068**	**0.900**	**0.880**	**0.064**	**0.838**	**0.731**	**0.064**	**0.926**	**0.887**	**0.044**	**0.830**	**0.773**	**0.105**	**0.938**	**0.903**	**0.041**
			7	layer4	layer8	0.830	0.734	0.070	0.902	0.879	0.064	0.835	0.723	0.066	0.924	0.881	0.045	0.831	0.770	0.105	0.937	0.899	0.042
			8	layer2	layer7	0.765	0.693	0.078	0.836	0.829	0.087	0.797	0.702	0.069	0.875	0.850	0.059	0.767	0.716	0.125	0.913	0.886	0.049
			9	layer3	layer7	0.798	0.725	0.071	0.874	0.861	0.073	0.818	0.721	0.064	0.903	0.875	0.050	0.794	0.739	0.115	0.928	0.899	0.044
			10	layer4	layer7	0.799	0.726	0.070	0.872	0.858	0.074	0.812	0.717	0.066	0.903	0.873	0.050	0.799	0.749	0.112	0.928	0.898	0.044

$1.32\%(F_\beta)$, $0.43\%(\mathcal{M})$ for MSRA-B and $0.42\%(E_\xi)$, $0.24\%(F_\beta)$, $0.22\%(\mathcal{M})$ for DUTS, showing that making each visual contrast cluster contribute fairly to the final saliency prediction may enrich the saliency representation and thus benefiting the final prediction. To verify the necessity of the proposed image weighting strategy, we perform experiments $4, 8, 12$ on the MSRA-B dataset and experiments $16, 20, 24$ on the DUTS dataset. The performances are further improved by a large margin, improving by $3.21\%(E_\xi)$, $1.46\%(F_\beta)$, $0.94\%(\mathcal{M})$ for MSRA-B and $0.51\%(E_\xi)$, $0.56\%(F_\beta)$, $0.38\%(\mathcal{M})$ for DUTS, suggesting that the IWS imposed on the main objective function forces the model to focus on those off-center objects, thus diminishing the misleading impact of the spatial distribution bias.

Network & Handcrafted Method Robustness. According to Tab. 2, by changing the backbone network $f(\cdot)$ from DRN-105 (experiments 1 to 12) to ResNet-50 (experiments 13 to 24), our framework still exhibits strong performance. Further, by replacing the training set from MSRA-B to DUTS, the performance on 6 datasets generally improves by a clear margin. This evidence validates that our proposed framework is effective and robust to the backbone network's choice. It also evidently show that our causal debiasing could improve the performance of individual handcrafted method, namely DSR, RBD, and MC, to be higher than or competitive to that of the SOTA method(s).

Contrast Feature Selection. To verify the selection of features for the construction of $\boldsymbol{C}_l$ and $\boldsymbol{C}_h$, we perform extensive experiments with DSR on MSRA-B and DUTS datasets using different levels of feature combinations. For ResNet-50, we choose res2, res3 as low-level visual contrast feature candidates, and res4,res5 as high-level visual contrast feature candidates; For DRN, we select layer2,layer3,layer4 as low-level feature candidates and layer7,layer8 as high-level feature candidates. According to Tab. 3, when training on DUTS dataset with ResNet-50, selecting res5 for high-level visual contrast representation performs better, indicating that semantically richer feature is crucial to obtain reliable representative high-level visual contrast. Besides, the choice of low-level visual contrast representation works equally well for both res2 and res3. When training on MSRA-B dataset with DRN, it shows a similar phenomenon for the choice of low-level and high-level visual contrast representations, that is choosing layer8 performs much better and choosing layer2, layer3, or layer4 performs almost equally well. This coincides with the study of neural network interpretability (Zeiler and Fergus 2014; Zhou et al. 2016), where the first three layers of neural networks contain rich low-level features and the last three layers contain rich high-level semantic features.

Grid Confounder Set Size & Design. Please refer to **the supplementary document** for more details.

Conclusion

This paper proposed a new causal debiasing framework to eradicate the detrimental contrast distribution bias and spatial distribution bias in USOD. Concretely, we proposed a causal graph to analyze the confounding effect of visual contrast distribution and identify that visual contrast distribution is a confounder which misleads the model training towards data-rich visual contrast clusters; then, we introduced a deconfounded training with causal intervention to make each visual contrast cluster contribute fairly. Further, we observed that the spatial object distribution biases the model training to focus on the center area of an image plane. To eradicate this bias, we proposed an image-level weighting strategy to weigh each image's importance softly. Extensive experiments have demonstrated our framework's effectiveness. Nevertheless, designing a better causal graph for USOD remains an open problem for future research.

Acknowledgements

This work was supported in part by the National Key Research and Development Program of China (No.2020YFC2003902), in part by the Guangdong Basic and Applied Basic Research Foundation under Grant No.2020B1515020048, in part by the National Natural Science Foundation of China under Grant No.61976250 and No.U1811463, and in part by the Guangdong Provincial Key Laboratory of Big Data Computing, the Chinese University of Hong Kong, Shenzhen.

References

Ahmed, O.; Trauble, F.; Goyal, A.; Neitz, A.; Wuthrich, M.; Bengio, Y.; Scholkopf, B.; and Bauer, S. 2021. CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning. In *ICLR*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *TPAMI*, 40: 834–848.

Chen, Z.; Xu, Q.; Cong, R.; and Huang, Q. 2020. Global Context-Aware Progressive Aggregation Network for Salient Object Detection. In *AAAI*.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *IJCAI*.

Feng, M.; Lu, H.; and Ding, E. 2019. Attentive Feedback Network for Boundary-Aware Salient Object Detection. *CVPR*, 1623–1632.

Goferman, S.; Zelnik-Manor, L.; and Tal, A. 2012. Context-Aware Saliency Detection. *TPAMI*.

Gupta, A.; Seal, A.; Prasad, M.; and Khanna, P. 2020. Salient Object Detection Techniques in Computer Vision—A Survey. *Entropy*, 22.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Itti, L.; and Koch, C. 2001. Computational modelling of visual attention. *Nature Reviews Neuroscience*.

Ji, Z.; Wang, H.; Han, J.; and Pang, Y. 2019. Saliency-Guided Attention Network for Image-Sentence Matching. *ICCV*.

Jiang, B.; Zhang, L.; Lu, H.; Yang, C.; and Yang, M.-H. 2013. Saliency Detection via Absorbing Markov Chain. *ICCV*.

Judd, T.; Ehinger, K. A.; Durand, F.; and Torralba, A. 2009. Learning to predict where humans look. *ICCV*.

Keith, K. A.; Jensen, D. D.; and O'Connor, B. 2020. Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates. In *ACL*.

Li, G.; Xie, Y.; and Lin, L. 2018. Weakly Supervised Salient Object Detection Using Image Labels. In *AAAI*.

Li, G.; and Yu, Y. 2015. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5455–5463.

Li, G.; and Yu, Y. 2016. Deep contrast learning for salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 478–487.

Li, X.; Lu, H.; Zhang, L.; Ruan, X.; and Yang, M.-H. 2013. Saliency Detection via Dense and Sparse Reconstruction. *ICCV*.

Liu, J.; Hou, Q.; Cheng, M.-M.; Feng, J.; and Jiang, J. 2019. A Simple Pooling-Based Design for Real-Time Salient Object Detection. *CVPR*, 3912–3921.

Liu, N.; Han, J.; and Yang, M.-H. 2018. PiCANet: Learning Pixel-Wise Contextual Attention for Saliency Detection. *CVPR*, 3089–3098.

Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; and Shum, H. 2007. Learning to Detect a Salient Object. *TPAMI*, 33: 353–367.

Liu, X.; Yin, D.; Feng, Y.; Wu, Y.; and Zhao, D. 2021. Everything Has a Cause: Leveraging Causal Inference in Legal Text Analysis. In *NAACL*.

Luo, Z.; Mishra, A.; Achkar, A.; Eichel, J. A.; Li, S.; and Jodoin, P.-M. 2017. Non-local Deep Features for Salient Object Detection. *CVPR*, 6593–6601.

Marchesotti, L.; Cifarelli, C.; and Csurka, G. 2009. A framework for visual saliency detection with applications to image thumbnailing. *ICCV*.

Nguyen, D. T.; Dax, M.; Mummadi, C. K.; Ngo, T.-P.-N.; Nguyen, T. H. P.; Lou, Z.; and Brox, T. 2019. DeepUSPS: Deep Robust Unsupervised Saliency Prediction With Self-Supervision. In *NeurIPS*.

Pang, Y.; Zhao, X.-Q.; Zhang, L.; and Lu, H. 2020. Multi-Scale Interactive Network for Salient Object Detection. *CVPR*, 9410–9419.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Pearl, J.; Glymour, M.; and Jewell, N. 2016. Causal Inference in Statistics: A Primer.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jägersand, M. 2019. BASNet: Boundary-Aware Salient Object Detection. *CVPR*, 7471–7481.

Reinagel, P.; and Zador, A. 1999. Natural scene statistics at the centre of gaze. *Network*.

Shelhamer, E.; Long, J.; and Darrell, T. 2017. Fully Convolutional Networks for Semantic Segmentation. *TPAMI*.

Shetty, R.; Fritz, M.; and Schiele, B. 2018. Adversarial Scene Editing: Automatic Object Removal from Weak Supervision. In *NeurIPS*.

Simakov, D.; Caspi, Y.; Shechtman, E.; and Irani, M. 2008. Summarizing visual data using bidirectional similarity. *CVPR*.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased Scene Graph Generation From Biased Training. In *CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *NeurIPS*, 5998–6008.

Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; and Ruan, X. 2017. Learning to Detect Salient Objects with Image-Level Supervision. *CVPR*, 3796–3805.

Wang, T.; Huang, J.; Zhang, H.; and Sun, Q. 2020. Visual Commonsense R-CNN. In *CVPR*.

Wang, T.; Zhang, L.; Wang, S.; Lu, H.; Yang, G.; Ruan, X.; and Borji, A. 2018. Detect Globally, Refine Locally: A Novel Approach to Saliency Detection. *CVPR*, 3127–3135.

Wei, J.; Wang, S.; and Huang, Q. 2020. F³Net: Fusion, Feedback and Focus for Salient Object Detection. In *AAAI*.

Wu, R.; Feng, M.; Guan, W.; Wang, D.; Lu, H.; and Ding, E. 2019. A Mutual Learning Method for Salient Object Detection With Intertwined Multi-Supervision. *CVPR*, 8142–8151.

Wu, Z.; Su, L.; and Huang, Q. 2019. Stacked Cross Refinement Network for Edge-Aware Salient Object Detection. *ICCV*, 7263–7272.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*.

Yu, F.; Koltun, V.; and Funkhouser, T. 2017. Dilated Residual Networks. In *CVPR*.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and Understanding Convolutional Networks. In *ECCV*.

Zhang, D.; Han, J.; and Zhang, Y. 2017. Supervision by Fusion: Towards Unsupervised Learning of Deep Salient Object Detector. *ICCV*.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.; and Sun, Q. 2020a. Causal Intervention for Weakly-Supervised Semantic Segmentation. In *NeurIPS*.

Zhang, J.; Dai, Y.; Zhang, T.; Harandi, M.; Barnes, N.; and Hartley, R. 2020b. Learning Saliency from Single Noisy Labelling: A Robust Model Fitting Perspective. *TPAMI*.

Zhang, J.; Xie, J.; and Barnes, N. 2020. Learning Noise-Aware Encoder-Decoder from Noisy Labels by Alternating Back-Propagation for Saliency Detection. In *ECCV*.

Zhang, J.; Zhang, T.; Dai, Y.; Harandi, M.; and Hartley, R. 2018. Deep Unsupervised Saliency Detection: A Multiple Noisy Labeling Perspective. *CVPR*.

Zhao, J.; Liu, J.; Fan, D.-P.; Cao, Y.; Yang, J.; and Cheng, M.-M. 2019a. EGNet: Edge Guidance Network for Salient Object Detection. *ICCV*, 8778–8787.

Zhao, K.; Gao, S.; Hou, Q.; Li, D.; and Cheng, M.-M. 2019b. Optimizing the F-Measure for Threshold-Free Salient Object Detection.

Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization.

Zhu, W.; Liang, S.; Wei, Y.; and Sun, J. 2014. Saliency Optimization from Robust Background Detection. *CVPR*.