Joint Semantic Mining for Weakly Supervised RGB-D Salient Object Detection

Anonymous Author(s)
Affiliation
Address
email

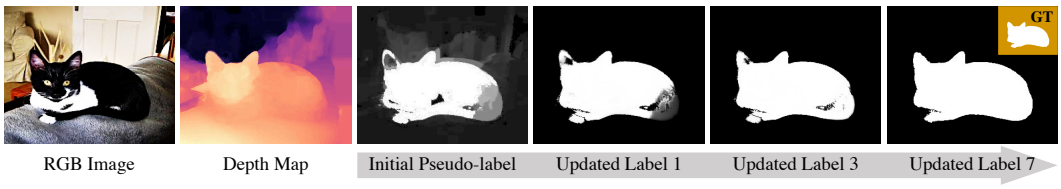
Abstract

Training saliency detection models with weak supervisions, *e.g.,* image-level tags or captions, is appealing as it removes the costly demand of per-pixel annotations. Despite the rapid progress of RGB-D saliency detection in fully-supervised setting, it however remains an unexplored territory when only weak supervision signals are available. This paper is set to tackle the problem of *weakly-supervised RGB-D salient object detection*. The key insight in this effort is the idea of maintaining per-pixel pseudo-labels with iterative refinements by reconciling the multimodal input signals in our joint semantic mining (JSM). Considering the large variations in the raw depth map and the lack of explicit pixel-level supervisions, we propose spatial semantic modeling (SSM) to capture saliency-specific depth cues from the raw depth and produce depth-refined pseudo-labels. Moreover, tags and captions are innovatively incorporated via a fill-in-the-blank training in our textual semantic modeling (TSM) to estimate the confidences of competing pseudo-labels. At test time, our model involves only a light-weight sub-network of the training pipeline, *i.e.*, it requires only an RGB image as input, thus allowing efficient inference. Extensive evaluations demonstrate the effectiveness of our approach under the weakly-supervised setting. Importantly, our method could also be adapted to work in both fully-supervised and unsupervised paradigms. In each of these scenarios, superior performance has been attained by our approach with comparing to the state-of-the-art dedicated methods. As a by-product, a *CapS* dataset is constructed by augmenting existing benchmark training set with additional image tags and captions. *The code & dataset will be publicly available.*

1 Introduction

As a fundamental computer vision task, salient object detection (SOD) aims at locating and segmenting visually distinctive objects in an image scene. It plays an important role in a variety of downstream applications including image retrieval [29], object recognition [51], and video analysis [56]. Recent progress in supervised RGB-D SOD [3, 4, 10, 32, 39] has demonstrated significant benefits of engaging depth information in saliency detection from complex scenes. The success of these fully-supervised methods, however, relies heavily on the large-scale, precise, pixel-level annotations, which are often laborious and time-consuming to acquire. On the other hand, an image usually comes with additional information in its meta-data such as tags and captions from users to describe the scene context and content, which may serve as cheap weak-supervision signals. These weak supervision signals are nonetheless noisy and have mixed qualities. Similar weak-supervision signals have been explored in RGB-image based SOD [55, 58, 34], where the noisy nature of these side information is unfortunately overlooked. To further complicate the matter, the lack of explicit pixel-level supervision brings new challenge to the RGB-D SOD task: the depth values from raw depth maps are often noisy and sometimes inconsistent. For example, in Fig. 1, similar depth values are shared by the cat and the underneath couch, making it difficult to discern the salient object from backgrounds. Without the explicit pixel-level supervision, existing cross-modal fusion strategies adopted by fully-supervised RGB-D methods [32, 46, 33] would simply fail.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.



Category:	Caption:	The masked Caption:
Cat	A black and white **cat** is laying on a gray couch.	A black and white <MASK> is laying on a gray couch.

Figure 1: Illustration of weakly-supervised RGB-D salient object detection. RGB and depth images, as well as weak supervision signals such as image-level tags and captions are exploited. Initial pseudo-label is generated by the handcrafted methods, which is then iteratively updated by our joint semantic mining pipeline. GT is ground-truth label for reference.

These observations motivate us to consider the new problem of *weakly-supervised RGB-D salient object detection*, which takes as input the RGB and depth images, as well as weak supervision signals such as image-level tags and captions, as illustrated in Fig. 1. By removing the demand for laborious per-pixel annotations, it also brings new challenges: 1) how to address the noisy nature of the weak supervision signals; 2) how to tackle the depth noise and inconsistency to facilitate proper separation of foreground and background regions.

This leads us to propose the use of pseudo-labels with iterative refinements in training: the pseudo-label provides internal pixel-level supervision signals, which is progressively updated by reconciling the multimodal input signals and the current information flow of the neural net, based on the previous pseudo-label. As illustrated in Fig. 2, this is realized by an interaction between two core modules, namely spatial semantic modeling (SSM) and joint semantic mining (JSM): SSM is designed to capture the saliency-specific depth semantics, to eliminate the background noises in the coarse saliency prediction, and to generate a depth-refined pseudo-label. This simple yet effective module is very generic, which could be easily plugged-in different setups, including unsupervised & fully-supervised scenarios; meanwhile, the JSM module is proposed to leverage depth semantics and weak supervision signals for attaining more reliable pseudo-labels. Specifically, a partial textual input, *i.e.,* image-level tag and caption with its salient word being masked, is fed into a dedicated textual semantic modeling or TSM to estimate the confidence scores of competing pseudo-labels, and to fill-in-the-blank. Intuitively, a semantically consistent pseudo-label should provide better context cues to reconstruct the salient word; while a closer guess of the masked word would indicate a better pseudo-label. The alternation between SSM and JSM modules is thus expected to give rise to more trustworthy pseudo-labels. At test time, it is then sufficient to take an input RGB image and activate a light-weight network to deliver its final prediction. That is, test time input involves only an RGB image, without the need of any depth map or image-level tags and captions. This drastically simplifies the input requirement and reduces the computation burden at deployment stage. Moreover, given the lack of training dataset for the weakly-supervised setting, we adapt existing RGB-D training dataset to annotate additional image-level tags and captions, which is referred to as the *CapS* dataset.

The main contributions of this paper are as follows. (1) A new problem of weakly-supervised RGB-D salient object detection is considered. In this regard, a *CapS* dataset is curated by augmenting the existing RGB-D SOD training dataset with image-level tagging and captioning annotations. (2) The key ingredient of our approach involves the production of pseudo-labels with iterative refinements, realized by iterative updates between two internal modules, SSM and JSM. Empirical experiments demonstrate the effectiveness of our approach in weakly-supervised setting. Moreover, (3) after proper adaptation of our approach in unsupervised and fully-supervised scenarios, superior performance is also observed when comparing to the respective state-of-the-art methods. (4) Our test time inference amounts to executing a light-weight saliency network: as illustrated in dotted line at Fig. 2, only an RGB image is used as its input, thus allows efficient and effective inference. *Our dataset and source code will be made publicly available*.

2 Related Work

RGB-D salient object detection [44, 74] has been an active field of research in the past few years, where the incorporation of depth cues has been demonstrated [19, 12, 13, 15, 17, 31, 37, 52, 70] to improve performance especially in complex scenes. Existing RGB-D methods aim to design effective feature fusion strategies for learning representative cross-modal features. Typically, Chen *et al.* [23, 7] employ two-stream CNNs-based models and perform fusion by adding or concatenating paired features at shallow or deep layers. Fu *et al.* [21] utilize a Siamese network to jointly learn RGB

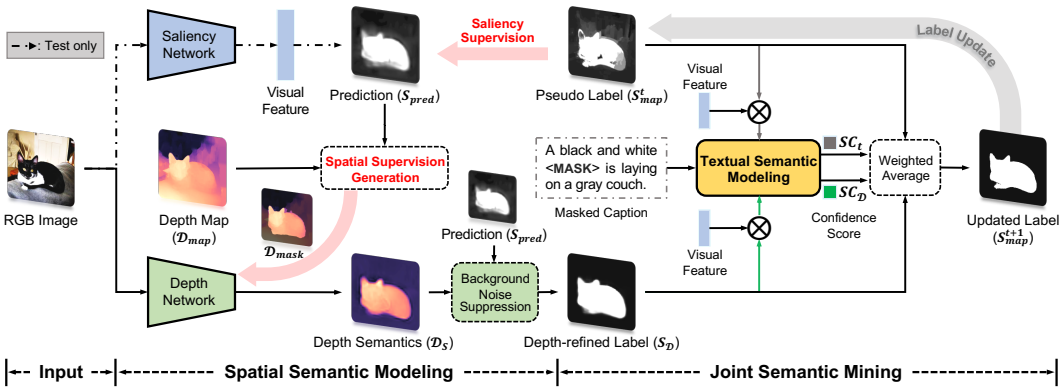


Figure 2: An overview of our approach. Its training pipeline consists of a saliency prediction network, a SSM (Sec. 3.2) to generate depth-refined pseudo-label, a TSM (Sec. 3.3) to estimate the confidences of different pseudo-labels, and a JSM (Sec. 3.4) to refine & update pseudo-label. Our testing process only involves activating a saliency network delineated in dotted lines. More details of SSM and TSM modules are illustrated in Fig. 3. The masked salient word is 'cat'.

and depth inputs for mining useful complementary features. To promote multi-modal interactions, Li *et al.* [33] design a cross-modal weighting strategy to encourage comprehensive interactions between RGB and depth information. However, those methods unfortunately rely on costly pixel-level annotations, which are tedious and time consuming to acquire. This motivates us to consider a weakly-supervised approach. In what follows, our focus will be mainly toward related weakly-supervised methods developed for saliency detection from RGB images, where the differences of our approach from existing methods would be clarified.

Instead of using costly pixel-level annotations, some recent efforts instead explore cheap alternatives such as image-level tags [34, 55, 59], image captions [58], scribble labels [57, 64, 71], and noisy pseudo-labels from handcrafted methods [41, 60, 61, 63, 65]. The pioneering work [55] leverages image-level tags or categories that could be augmented onto existing large-scale dataset at low-cost. The same scenario is also considered by Li *et al.* [34], where a composite pipeline combining graphical model with CNNs is designed. The trained network however tends to highlight the most discriminative region instead of the intended salient object out of the scene due to the sparse image-level supervisions. Image captions are examined by Zeng *et al.* [58] as supervision input; in their work image classification network and caption generation network are jointly trained to obtain pseudo-labels, which achieves descent performance. Scribble is another type of weak supervision signal, where a tiny fraction of image pixels are labeled by users as being foreground or background. Due to the annotation sparsity, object structure and details cannot be easily inferred. Zhang *et al.* [64] introduce a gated structure-aware loss as well as an auxiliary edge detection network to uncover the complete object. Meanwhile, Yu *et al.* [57] explore self-consistency among multi-scale outputs and design a local coherence loss to propagate the labels to unlabeled regions based on image features, thus enabling the detection of objects with smooth textures. However, existing weakly supervised saliency methods are solely based on RGB image. Unlike the prevalence of fully-supervised RGB-D SOD, it has never been considered the incorporation of depth cues in existing literature.

This leads us to address this problem in the presence of image tags and captions as weak supervision signals. Different from existing methods [34, 26, 55, 58] that train image classification networks or caption generation networks to delineate potential salient regions, masked salient word in captions are to be reconstructed in our work by leveraging visual saliency features. This is then used to estimate the confidence scores of pseudo-labels.

3 Methodology

3.1 The Overall Architecture

An overview of our approach is illustrated in Fig. 2. It consists of a saliency network responsible for saliency prediction, a spatial (depth) semantic modeling (SSM) to generate depth-refined pseudo-label, a textual (caption) semantic modeling (TSM) to estimate the confidences of different pseudo-labels, and a joint semantic mining (JSM) strategy to refine & update pseudo-label. Overall our pipeline aims to gradually improve the quality of noisy pseudo-labels by jointly mining the useful spatial semantics from the depth map and textual semantics from the tags and captions, to produce more trustworthy supervision signals, which in turns results in better training of the saliency network.

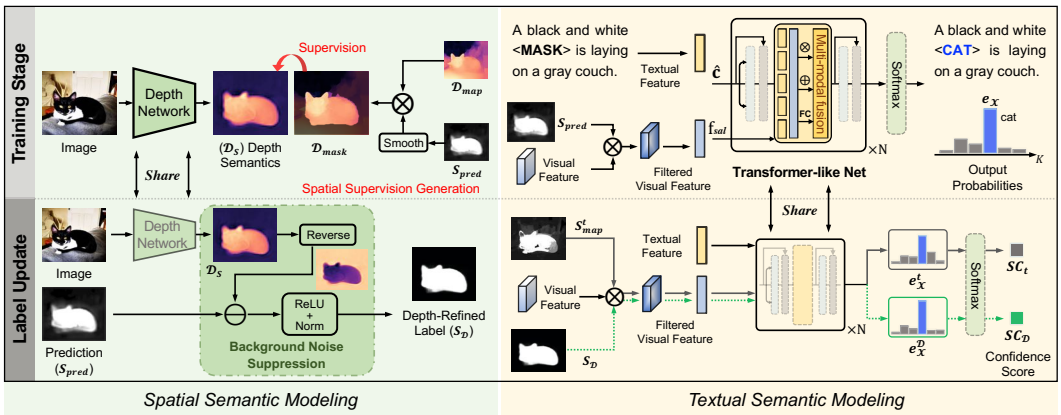


Figure 3: The detailed architecture of the proposed SSM and TSM. The upper shows their training processes, and the bottom illustrates the way of using them to perform label update.

Specifically, the widely-used DeepLabV3+ [46] architecture is adopted in both saliency network and depth network. Initial supervision signal for the saliency network is provided by traditional handcrafted methods. The predicted saliency map, together with the raw depth map, are processed to generate the saliency-guided spatial supervision signal for the depth network. Then the predicted saliency-oriented depth semantics is utilized to eliminate background noises (non-salient regions) in coarse prediction, and to generate a depth-refined pseudo-label. This is followed by our JSM strategy, which takes in the image-level tags and captions through TSM to estimate the confidence scores of pseudo-labels; updated pseudo-label is then formed based on the confidence-weighted depth-refined pseudo-label and current pseudo-label, which provides more trustworthy supervision signal for the saliency network. Note our test time inference involves only the black dashed portion in Fig. 2, which takes as input only the RGB image, thus enables efficient saliency prediction.

3.2 Saliency-oriented Spatial (Depth) Semantic Modeling

Initial pseudo-labels are generated by traditional saliency models, which often contain excessive noise. As illustrated in Fig. 3, our spatial semantic modeling (SSM) is to produce a more reliable depth-refined pseudo-label, achieved by explicitly capturing saliency-specific depth semantics from the depth map to eliminate possible background noise in the coarse saliency prediction.

Concretely, during training, we first generate a saliency-guided depth mask $\mathcal{D}_{mask}$ by multiplying the rough saliency prediction $\mathcal{S}_{pred}$ with the raw depth map $\mathcal{D}_{map}$ in a spatial attention manner. Here, a Gaussian smooth operation is applied to smooth the predicted saliency area, to effectively perceive and capture more saliency areas from depth. The procedure is formulated as:

$$\mathcal{D}_{mask} = \Omega_{max}(\mathcal{F}_{gauss}(\mathcal{S}_{pred}, k), \mathcal{S}_{pred}) \otimes \mathcal{D}_{map}, \tag{1}$$

where $\mathcal{F}_{gauss}(\cdot, k)$ indicates a convolution operation with Gaussian kernel k and zero bias; $\Omega_{max}(\cdot,\cdot)$ is a maximum function to preserve the higher values between the smoothed and the original maps. $\otimes$ is element-wise multiplication. In this paper, the size and standard deviation of kernel k are learnable through the model training procedure and are initialized with values 32 and 4, respectively.

After obtaining $\mathcal{D}_{mask}$, a depth network is trained to learn the saliency-specific depth semantics $\mathcal{D}_\mathcal{S}$, using the mean square error (MSE) loss function. The internal inspection evidences in Fig. 4 suggest that $\mathcal{D}_\mathcal{S}$ (depth semantics, 6th column) is able to capture discriminative saliency cues from the raw depth map under the supervision of $\mathcal{D}_{mask}$ (5th column).

In addition, the learned $\mathcal{D}_\mathcal{S}$ can be further processed to generate the depth-refined pseudo-labels. This procedure is only employed when performing pseudo-label update. Specifically, we feed $\mathcal{D}_\mathcal{S}$ into an Background Noise Suppression block to help eliminate the background noises (non-salient regions) in the coarse prediction $\mathcal{S}_{pred}$. In this block, a reverse operation is first employed on $\mathcal{D}_\mathcal{S}$ to highlight background regions by $1 - \mathcal{D}_\mathcal{S}$. This is followed by the pixel-wise subtraction to suppress the non-salient negative responses in $\mathcal{S}_{pred}$. Finally, a ReLU function and a normalization procedure are adopted to suppress negative numbers and normalize the result to the range of [0, 1]. This procedure of obtaining the depth-refined pseudo-label $\mathcal{S}_\mathcal{D}$ could be expressed by

$$\mathcal{S}^{i,j}_\mathcal{D} = \frac{\mathcal{S}^{i,j}_d - min(\mathcal{S}_d)}{max(\mathcal{S}_d) - min(\mathcal{S}_d) + \varepsilon}, \ \mathcal{S}_d = \mathrm{ReLU}(\mathcal{S}_{pred} - \lambda_d(1 - \mathcal{D}_\mathcal{S})), \ i \in [1, H], j \in [1, W]. \tag{2}$$

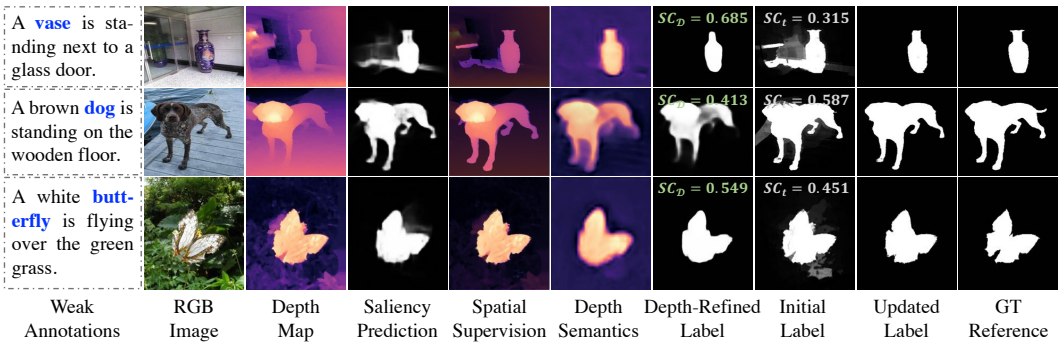


Figure 4: Step-by-step inspections of the internal processes of our approach. The GT is for reference.

Here $\lambda_d \in [0, 1]$ is a constant to control the degree of the subtracted background noises and avoid negatively suppressing salient regions that have relatively low saliency scores in $\mathcal{S}_{\mathcal{D}}$. Throughout experiments, λ_d is empirically set to 0.5, and ε to 1e-5. H and W are the height and width of the input image, respectively.

The pseudo-label refinement dynamics are visually inspected in Fig. 4, while Fig. 8 presents the corresponding quantitative error analysis over iterations. Empirical evidence suggests as the training proceeds, quality of the pseudo-label is significantly improved. Moreover, at this stage, we can directly utilize the depth-refined label $\mathcal{S}_{\mathcal{D}}$ to update pseudo-labels (as in Fig. 5 (b)). This can be referred as unsupervised RGB-D SOD since no weak labels are used.

3.3 Saliency-oriented Textual (Caption) Semantic Modeling

Previously, the mainstream use of weak labels is to train a classification network or a caption generation network, where the by-product attention maps or Class Activation Maps [73] are leveraged to determine the potential salient regions [58, 55]. It is very different in our textual semantic modeling (TSM), where the main focus is to leverage side information (*i.e.*, image-level tags and captions) to facilitate the production of reliable training signals. Inspired by the recent success of masked language models [16], captions with missing keywords are used as input, with the expectation of the complete text being reconstructed as output. In the proposed TSM, innovatively taking as input partial text with salient word being masked, as well as the saliency-filtered visual features, our TSM is to output the reconstructed text in a fill-in-the-blank manner and to estimate the confidence scores of competing pseudo-labels. The intuition is, a semantically matching pseudo-label could provide better context cues to reconstruct the masked salient word; meanwhile, a closer guess of the masked word would indicate a better pseudo-label.

Formally, for each training data, the weak labels contain caption description $\mathbf{c} = \{\mathbf{c}_i\}_{i=1}^{n_c}$, image-level category $k \in \{1, .., K\}$, and the position $\mathcal{X}$ of the salient word (object) in the caption, where n_c is the word number of the caption. Let $\mathbf{c}_i \in \mathbb{R}^{d\times 1}$ be the word embedding of the i-th word in the caption, K the total number of salient categories, and $\mathcal{X}$ an integral number. As shown in Fig. 3, during training, the input is a masked version of caption $\hat{\mathbf{c}} \in \mathbb{R}^{d\times n_c}$ where the salient word $\mathbf{c}_{\mathcal{X}}$ in caption $\mathbf{c}$ is masked with a special symbol. In order to reconstruct the masked salient word, we filter the visual feature from the saliency network by multiplying it with the learned saliency cues $\mathcal{S}_{pred}$. We then obtain the saliency-filtered visual feature, and transform it to a feature vector $\mathbf{f}_{sal} \in \mathbb{R}^{d\times 1}$ for subsequent cross-modal fusion using a pooling operation and a fully-connected (FC) layer.

The center component of the TSM module is a transformer-like network. Based on original Transformer [54], we add a multi-modal fusion sub-layer into network. Three parallel operations are used to promote sufficient cross-modal feature interactions: element-wise multiplication, element-wise addition, and concatenation followed by FC. The three outputs are concatenated and followed by a FC to change the feature dimension. Note the cross-modal fusion operation is performed word-wise. Through the textual network, we can obtain the energy vector $\mathbf{e}_{\mathcal{X}} \in \mathbb{R}^{K\times 1}$ of the masked salient word, which is computed over all categories by a fully-connected layer and softmax function. For each training sample, the training objective loss for the textual network is $-\log(\mathbf{e}_{\mathcal{X}}[k])$, where $\mathbf{e}_{\mathcal{X}}[k]$ is the output probability of salient category k. Detailed structure for our TSM can be accessed in the supplementary material. Once trained, the TSM module is then used to estimate the confidence scores of pseudo-labels as described in Sec. 3.4.

3.4 Joint Semantic Mining for Label Updating

The label updating operation using joint semantic mining is iteratively conducted every τ epochs (as a training round) during training, *i.e.*, the granularity of label updating. The choices are explored in the ablations of Sec. 4.3, where $\tau = 5$ is shown to work best empirically. In terms of training, the saliency network, the SSM module, and the TSM module are independently trained.

At the end of each training round, our JSM strategy performs the label updating operation to generate an up-to-date pseudo-label for each training image. Now define the pseudo-label for the saliency network in current iteration as $\mathcal{S}^t_{map}$. As shown in the label update phase of Fig. 3, the SSM module is engaged to generate the depth-refined pseudo-label $\mathcal{S}_\mathcal{D}$; It is passed to the TSM module, where the $\mathcal{S}^t_{map}$ and $\mathcal{S}_\mathcal{D}$ are taken as saliency attention maps to filter visual features, respectively. The TSM module infers the energy vectors $\mathbf{e}^t_\mathcal{X}$ and $\mathbf{e}^\mathcal{D}_\mathcal{X}$ for $\mathcal{S}^t_{map}$ and $\mathcal{S}_\mathcal{D}$, respectively. Thus their confidence scores $\mathcal{SC}_t$ and $\mathcal{SC}_\mathcal{D}$ can be calculated by (taking $\mathcal{SC}_t$ as an example):

$$\mathcal{SC}_t = \frac{\exp(\mathbf{e}^t_\mathcal{X}[k])}{\exp(\mathbf{e}^t_\mathcal{X}[k]) + \exp(\mathbf{e}^\mathcal{D}_\mathcal{X}[k])}. \tag{3}$$

The updated label is the weighted average of S^t_{map} and $\mathcal{S}_\mathcal{D}$: $\mathcal{S}^{t+1}_{map} = \mathcal{SC}_t \times \mathcal{S}^t_{map} + \mathcal{SC}_\mathcal{D} \times \mathcal{S}_\mathcal{D}$. A fully-connected conditional random field [25] is applied to generate the final updated label which could provide more trustworthy supervision signal to train the saliency network.

3.5 The CapS Dataset

To train the weakly-supervised RGB-D SOD network, we relabel two widely-used RGB-D saliency datasets, NJUD [27] and NLPR [44], which contain a total of 2,185 training images. We provide various image-level weak annotations: categories, captions, and position of the salient word in each of the captions. The annotation process is conducted semi-automatically: the NeuralTalk2 [28] toolkit is utilized to automatically generate image captions, which are then manually checked, with unreasonable cases corrected. On average, there are 8.9 words in each caption. This is followed by categorically tagging each of the images, each corresponds to a salient category; this is different from the traditional image classification dataset ImageNet [30]. We summarize the 100 categories in the *CapS*. The positions of the salient words are first automatically identified through localizing the image categories in the captions which is subsequently followed by manual identification. Detailed statistics and examples of our in-house *CapS* dataset are relegated to the supplementary material.

4 Experiments

4.1 Datasets and Evaluation Metrics

Empirical evaluations are conducted over four large-scale RGB-D SOD benchmark datasets, including NJUD [27] with 1,985 RGB-D paired images, NLPR [44] with 1,000 samples, STERE [42] with 1,000 stereoscopic images, and DUTLF-Depth [46] with 1,200 RGB-D data. In train vs. test splits of the datasets, the popular setup of [20, 21, 62] is followed for a fair comparison. Training set consists of 1,485 samples from NJUD and 700 samples from NLPR. The remaining images in these datasets and other public test sets are reserved for testing purposes throughout the experiments. Four widely-used metrics are adopted for quantitative evaluation: they are E-measure (E_ξ) [18], F-measure (F_β) [1], weighted F-measure (F^w_β) [40],and mean absolute error ($\mathcal{M}$) [2].

4.2 Implementation Details and Setups

We use DeepLabV3+ [8] with ResNet-50 [24] pre-trained on ImageNet as backbone, for both saliency network and depth network. For each word in the caption, we extract word embedding with dimension $d = 300$ using the pretrained Glove [45] word2vec network. The maximum caption length is set to 20. For the transformer-like net in TSM, we set the hidden state to 256, the number of layers to 3, and the number of head to 4. The model is optimized by Adam with batch size of 10, and the learning rate is set to 1×10^{-4}. During training, we use the standard BCE loss to train the saliency network. Each image is uniformly resized to 352×352 and is performed by randomly rotating and cropping to avoid potential overfitting. Our network is trained in an end-to-end manner and converges around 40 epochs. The code is implemented in Pytorch toolbox on a PC with a single Tesla P40 GPU. Initial pseudo labels are generated by the handcrafted method [77], which is freely available for annotations.

4.3 Empirical Analysis of our Pipeline

Here we focus on evaluating the contribution of each component in our pipeline, and examining the internal performance of the pseudo-labels at different stages in training.

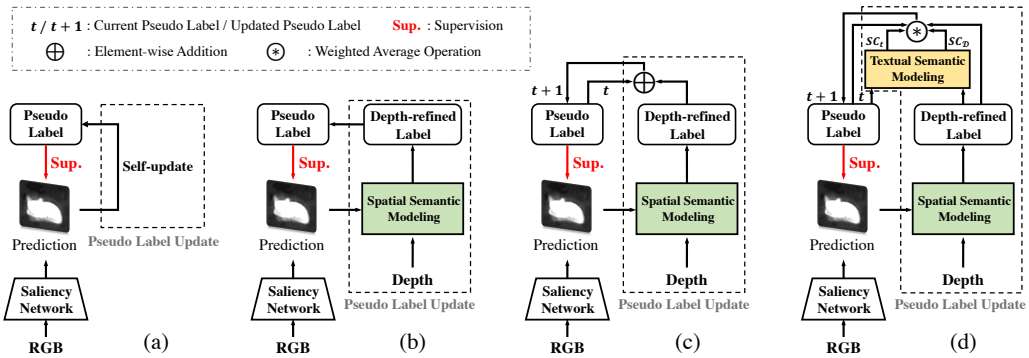


Figure 5: Diagrams of various label updating strategies used in our ablation study: (a) self-updating strategy, (b) SSM updating strategy, (c) historical moving average with equal weights, and finally (d) our JSM strategy.

Table 1: Ablation study of our pipeline. ↑ (↓) denote performance gains (relative to backbone).

Model Setups	NJUD [27]			NLPR [44]			STERE [42]			DUTLF-Depth [46]		
	F_β^w	F_β	$\mathcal{M}$	F_β^w	F_β	$\mathcal{M}$	F_β^w	F_β	$\mathcal{M}$	F_β^w	F_β	$\mathcal{M}$
Backbone trained on Pseudo Labels (*i.e.,* '$\mathcal{B}$')	.510	.627	.186	.479	.570	.126	.555	.666	.158	.512	.644	.167
'$\mathcal{B}$' trained on Pseudo Labels with CRF	.546	.642	.165	.585	.624	.094	.601	.684	.135	.568	.670	.140
'$\mathcal{B}$' with Self-updating Strategy	.571	.673	.154	.607	.651	.087	.643	.708	.123	.616	.697	.130
'$\mathcal{B}$' with Spatial Semantic Modeling	.576	.689	.147	.638	.698	.073	.654	.738	.111	.629	.741	.116
	↑13%	↑10%	↓21%	↑33%	↑22%	↓42%	↑18%	↑11%	↓30%	↑23%	↑15%	↓31%
'$\mathcal{B}$' with SSM and Historical Moving Average	.588	.698	.141	.664	.722	.068	.674	.753	.105	.644	.750	.113
'$\mathcal{B}$' with Joint Semantic Mining (*i.e.,* **Ours**)	**.599**	**.717**	**.133**	**.692**	**.770**	**.060**	**.688**	**.778**	**.095**	**.698**	**.797**	**.093**
	↑17%	↑14%	↓28%	↑44%	↑35%	↓52%	↑24%	↑17%	↓40%	↑36%	↑24%	↓44%

Ablation analysis. We present in Table 1 the ablation results of our pipeline on four benchmarks. To start with, we consider the backbone as the saliency network trained with initial pseudo labels. As our proposed SSM and TSM are gradually incorporated into the backbone to generate the depth-refined pseudo-labels and estimate their confidence scores for label updating, noticeable performance gains are consistently achieved in all datasets. The SSM significantly reduces the MAE metric and increases the F-measure score by 31% and 14.5% on average in four datasets, respectively. The TSM further boosts the saliency detection performance to a higher level where a significant amount of performance gains with 41% and 22.5% are finally achieved on MAE and F-measure metrics.

To further demonstrate the effectiveness of our SSM in exploiting the depth semantics to refine pseudo-labels, we retrain the saliency network using the self-updating strategy as in Fig. 5 (a). In this strategy, the saliency prediction with CRF are directly utilized to update pseudo-labels, at the end of each training round. As shown in Table 1 (3rd row *vs.* 4th row), when excluding saliency-oriented depth semantics captured by the SSM, the performance of model degrades greatly. This indicates our SSM can effectively suppress background noise and providing reliable training labels. Furthermore, we replace TSM with a heuristic historical moving average strategy where the pseudo-labels are assigned with equal weights as in Fig. 5 (c). Table 1 shows it achieves better performance than the backbone using SSM due to the consideration of historical information, but it is consistently inferior compared to our pipeline with the TSM module. These results suggest that our TSM can better estimate the confidence or quality of pseudo-labels and generate the trustworthy supervision signals. Meanwhile, we also provide the internal inspections of our approach in Fig. 4, in terms of the generation of pseudo-labels and their corresponding confidence scores, for better understanding.

In addition, we further discuss the effect of different update intervals in our pipeline when performing label updating. As listed in Table. 2, the larger or smaller update interval leads to inferior performance due to the insufficient or excessive learning of models.

Table 2: Parameter analysis of the update interval τ (epoch) in the JSM.

Interval (τ)	NJUD [27]		NLPR [44]		STERE [42]	
	F_β	$\mathcal{M}$	F_β	$\mathcal{M}$	F_β	$\mathcal{M}$
$\tau = 1$	.703	.142	.766	.063	.769	.098
$\tau = 5$	.717	.133	.770	.060	.778	.095
$\tau = 10$	.695	.138	.763	.065	.758	.101

Analysis of pseudo labels. We analyze the evolutions of pseudo-labels in our training process. Here the quality of pseudo-labels is measured using the ground-truth labels of the training set (for evaluation only). Visual evidences in Fig. 7 show that the quality of pseudo-labels is gradually improved as our JSM is performed. We can see that the initial pseudo-labels unfortunately tend to miss important salient parts as well as fine-grained details. By adopting our proposed joint semantic mining for label updating, the missing parts could be gradually retrieved, with the object silhouette also being refined. The final pseudo-label is closest to the true label, which could provide more

Table 3: Quantitative comparison with weakly-supervised and unsupervised saliency models. Note RGB-based methods are specifically marked by ‡. $\mathcal{D}_\mathcal{S}$ and $\mathcal{T}_\mathcal{S}$ represent the spatial semantic modeling and textual semantic modeling, respectively. 'Un' means unsupervised learning. 'Cls' is SOD with class label. 'Cap' represents using weak supervision signals, with both class label and image caption.

*	Sup.	NJUD [27]				NLPR [44]				STERE [42]				DUTLF-Depth [46]			
		E_ξ	F_β^w	F_β	$\mathcal{M}$	E_ξ	F_β^w	F_β	$\mathcal{M}$	E_ξ	F_β^w	F_β	$\mathcal{M}$	E_ξ	F_β^w	F_β	$\mathcal{M}$
RBD‡ [78]	Un	.684	.387	.556	.256	.765	.388	.590	.211	.730	.443	.610	.223	.733	.447	.619	.222
MST‡ [53]	Un	.670	.291	.436	.281	.762	.257	.491	.199	.681	.312	.447	.269	.678	.254	.401	.279
BSCA‡ [48]	Un	.756	.446	.623	.216	.745	.376	.554	.178	.803	.497	.676	.179	.808	.479	.682	.181
DSR‡ [35]	Un	.739	.436	.594	.196	.757	.451	.545	.120	.785	.486	.645	.165	.797	.478	.640	.164
ACSD [27]	Un	.760	.448	.676	.198	.751	.327	.547	.171	.793	.425	.661	.200	.250	.210	.188	.668
DES [11]	Un	.421	.241	.165	.448	.735	.259	.583	.301	.579	.281	.594	.295	.733	.386	.668	.280
LHM [44]	Un	.722	.311	.625	.201	.772	.320	.520	.119	.484	.379	.703	.172	.767	.350	.659	.174
GP [50]	Un	.730	.323	.666	.204	.813	.347	.670	.144	.785	.371	.710	.182	-	-	-	-
CDB [36]	Un	.752	.408	.650	.200	.810	.388	.618	.108	.808	.436	.713	.166	-	-	-	-
SE [22]	Un	.779	.518	.681	.164	.853	.578	.695	.085	.825	.546	.732	.143	.730	.339	.474	.196
DCMC [14]	Un	.776	.506	.675	.167	.684	.265	.328	.196	.655	.551	.727	.148	.712	.290	.406	.243
MB [76]	Un	.643	.369	.492	.202	.814	.574	.637	.089	.693	.455	.572	.178	.691	.464	.577	.156
CDCP [77]	Un	.751	.522	.618	.181	.785	.512	.591	.114	.797	.596	.666	.149	.794	.530	.633	.159
USD‡ [65]	Un	.768	.565	.630	.163	.786	.536	.580	.119	.796	.572	.670	.146	.795	.545	.650	.157
DUSPS‡ [41]	Un	.771	.576	.647	.159	.809	.622	.639	.088	.806	.632	.682	.124	.798	.573	.654	.149
Ours ($\mathcal{D}_\mathcal{S}$)	Un	**.781**	**.576**	**.689**	**.147**	**.857**	**.638**	**.698**	**.073**	**.838**	**.654**	**.738**	**.111**	**.845**	**.629**	**.741**	**.116**
WSS‡ [55]	Cls	.760	.566	.693	.149	.856	.651	.712	.077	.831	.664	.732	.115	.843	.634	.743	.125
MSW‡ [58]	Cap	.773	.580	.704	.141	.869	.659	.736	.071	.836	.675	.760	.103	.863	.678	.789	.105
Ours ($\mathcal{D}_\mathcal{S}$&$\mathcal{T}_\mathcal{S}$)	Cap	**.788**	**.599**	**.717**	**.133**	**.888**	**.692**	**.770**	**.060**	**.852**	**.688**	**.778**	**.095**	**.870**	**.698**	**.797**	**.093**

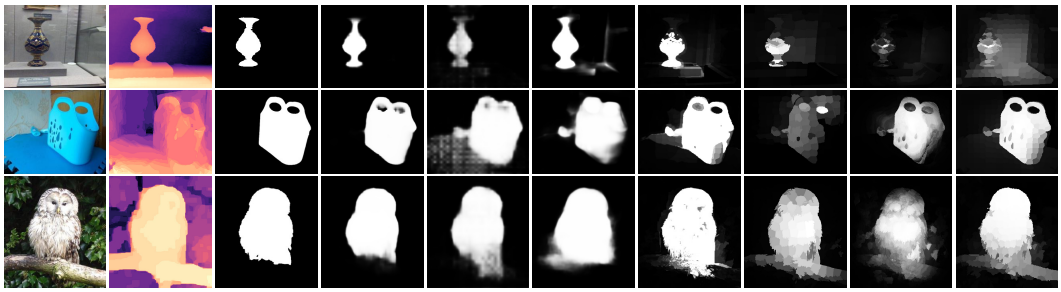


Image Depth GT Ours MSW DUSPS CDCP DCMC DSR BSCA

Figure 6: Visual comparison of weakly-supervised and unsupervised saliency models.

reliable guiding signal for training the saliency network. Moreover, we present in Fig. 8 the error reduction curves of the updated pseudo-labels with our full pipeline (blue line) and SSM only (red line), respectively. It can be seen that only SSM is able to improve the quality of pseudo-labels by exploiting the useful depth semantics to refine pseudo-labels. Our JSM further boosts the performance by leveraging the textual semantics to integrate reliable pseudo-labels.

4.4 Model Performance

We quantitatively evaluate the performance of our approach in Table 3, with visual results shown in Fig. 6. Since our approach is the first work for weakly-supervised RGB-D saliency detection, we show the results of two state-of-the-art RGB-based weakly-supervised methods (WSS [55] and MSW [58]) for reference. To make a fair comparison, we fine-tune them on the same training set using their published code with default setups. These results show that the effectiveness of our proposed method. Furthermore, our SSM and TSM in our joint semantic mining framework do not introduce any additional inference cost since they only participate in the training procedure to provide more reliable supervisory signals for saliency network. This design leads to a light-weight network that works both efficiently and effectively. As shown in Table 4, our method runs fastest among RGB-D methods, and also on par with the most efficient RGB-based methods.

Table 4: Inference time of different unsupervised and weakly-supervised saliency models. The RGB-based methods are specifically marked by ‡.

*	LHM [44]	DES [11]	GP [50]	CDCP [77]	DCMC [14]	SE [22]	ACSD [27]	CDB [36]	BSCA‡ [48]	RBD‡ [78]	MST‡ [53]	DSR‡ [35]	DUSPS‡ [41]	MSW‡ [58]	**Ours**
Inference Time (s)	2.13	7.79	12.98	5.7	1.2	1.57	0.718	0.6	2.665	0.1893	0.0302	0.3758	0.0292	0.0267	0.0286

4.5 Generalization Analysis

Adapting to unsupervised setting. Our approach can be adapted to unsupervised setting by using the architecture illustrated in Fig. 5 (b), where only the depth semantics are mined to refine pseudo-labels without weak labels. The quantitative results are presented in Table 3. Since existing unsupervised RGB-D methods are all based on handcrafted feature representations, we extend the comparison

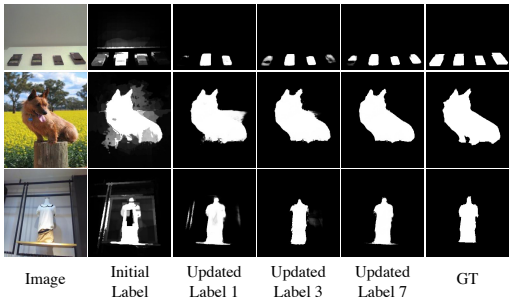


Figure 7: Visualization of updated labels.

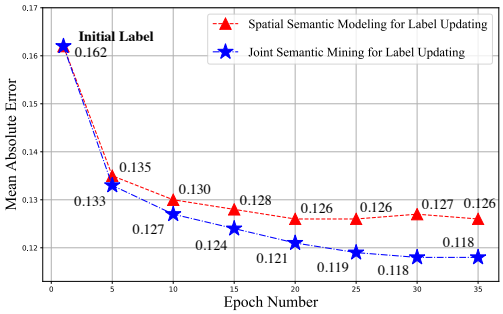


Figure 8: Error reduction plot of the updated pseudo-labels over iterations.

Table 5: Ablation analysis of the supervised variant of our approach, where the human annotations (*i.e.,* ground truths) are used to train SOD models.

Model Setups (fully)	NJUD [27]			NLPR [44]			STERE [42]		
	F_β^w	F_β	$\mathcal{M}$	F_β^w	F_β	$\mathcal{M}$	F_β^w	F_β	$\mathcal{M}$
Backbone (*i.e.,* '$\mathcal{B}$')	.864	.871	.046	.865	.863	.028	.858	.869	.045
'$\mathcal{B}$' + SSM (Ours)	.885	.906	.038	.892	.905	.022	.876	.896	.039

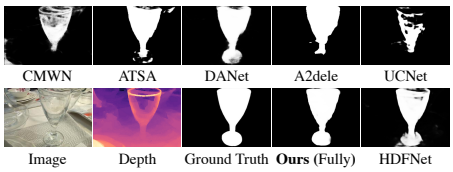


Figure 9: Visual comparison of fully supervised RGB-D SOD methods.

list to also include 2 recent RGB methods (USD [65] and DUSPS [41]) that engage the recent deep learning representations. For fair comparisons, we retrain them using their released implementations and default settings on the same training set. Numerical results show that our unsupervised variant clearly outperforms other unsupervised SOD methods in both RGB-D and RGB only scenarios.

Adapting to fully-supervised setting. Apart from the adaption to unsupervised setting, a variant of our approach can also be applied to fully-supervised RGB-D SOD scenario. This is achieved by modifying the generation of saliency-guided depth mask as $\mathcal{D}_{mask} = \mathcal{S}_{GT} \otimes \mathcal{D}_{map}$ in Eq. 1, with $\mathcal{S}_{GT}$ being the ground-truth annotations. The saliency network and depth network are trained by $\mathcal{S}_{GT}$ and $\mathcal{D}_{mask}$, respectively. Then the background noise suppression block in SSM module is applied to obtain the final saliency. As ablated in Table 5, our fully-supervised variant achieves consistent performance improvement compared to the backbone trained on $\mathcal{S}_{GT}$. In addition, our results compare favorably with those of 21 fully-supervised RGB-D saliency models, as quantitatively demonstrated in Table 6, and qualitatively illustrated in Fig. 9. Notice that, different from existing fully-supervised RGB-D SOD methods in designing the complicated cross-modal feature interaction strategy, our method directly exploits the learned depth semantics to promote saliency accuracy, which is simple yet effective and brand new for this field.

Table 6: Quantitative results of fully supervised RGB-D SOD methods.

Method	NJUD [27]				NLPR [44]				STERE [42]				DUTLF-Depth [46]			
	E_ξ	F_β^w	F_β	$\mathcal{M}$	E_ξ	F_β^w	F_β	$\mathcal{M}$	E_ξ	F_β^w	F_β	$\mathcal{M}$	E_ξ	F_β^w	F_β	$\mathcal{M}$
CTMF *ICyb'17* [23]	.864	.732	.788	.085	.869	.691	.723	.056	.841	.747	.771	.086	.884	.690	.792	.097
DF *TIP'17* [49]	.818	.552	.744	.151	.838	.524	.682	.099	.691	.596	.742	.141	.842	.542	.748	.145
PCA *CVPR'18* [5]	.896	.811	.844	.059	.916	.772	.794	.044	.887	.801	.826	.064	.858	.696	.760	.100
TANet *TIP'19* [6]	.893	.812	.844	.061	.916	.789	.795	.041	.893	.804	.835	.060	.866	.712	.779	.093
PDNet *ICME'19* [75]	.890	.798	.832	.062	.876	.659	.740	.064	.880	.799	.813	.071	.861	.650	.757	.112
MMCI *PR'19* [7]	.878	.749	.813	.079	.871	.688	.729	.059	.873	.757	.829	.068	.855	.636	.753	.113
CPFP *CVPR'19* [69]	.895	.837	.850	.053	.924	.820	.822	.036	.912	.808	.830	.051	.814	.644	.736	.099
DMRA *ICCV'19* [46]	.908	.853	.872	.051	.942	.845	.855	.031	.923	.841	.876	.049	.927	.858	.883	.048
SSF *CVPR'20* [67]	.913	.871	.886	.043	.949	.874	.875	.026	.921	.850	.867	.046	.946	.894	.914	.034
A2dele *CVPR'20* [47]	.897	.851	.874	.051	.945	.867	.878	.028	.915	.855	.874	.044	.924	.864	.890	.043
JL-DCF *CVPR'20* [21]	-	-	-	-	.954	.882	.878	**.022**	.919	.857	.869	.040	.931	.863	.883	.043
S2MA *CVPR'20* [38]	-	-	-	-	.938	.852	.853	.030	.907	.825	.855	.051	.921	.861	.866	.044
UCNet *CVPR'20* [62]	-	-	-	-	.953	.878	.890	.025	.922	.867	.885	**.039**	.903	.821	.856	.056
PGAR *ECCV'20* [9]	.915	.871	.893	.042	.955	.881	.885	.024	.919	.856	.880	.041	.944	.889	.914	.035
D3Net *NNLS'20* [19]	.913	.860	.863	.047	.943	.854	.857	.030	.920	.845	.855	.046	.847	.668	.756	.097
CMWN *ECCV'20* [33]	.910	.855	.878	.047	.940	.856	.859	.029	.917	.847	.869	.043	.916	.831	.866	.056
BBSNet *ECCV'20* [20]	.924	.884	.902	**.035**	.952	.879	.882	.023	.925	.858	.885	.041	.833	.663	.774	.120
DANet *ECCV'20* [72]	-	-	-	-	.949	.858	.871	.028	.914	.830	.858	.047	.925	.847	.884	.047
FRDT *ACMM'20* [68]	.917	.862	.879	.048	.946	.863	.868	.029	.925	.858	.872	.042	.941	.878	.902	.039
ATSA *ECCV'20* [66]	.921	.883	.893	.040	.945	.867	.876	.028	.919	.866	.874	.040	.947	.901	.918	.032
HDFNet *ECCV'20* [43]	.915	.879	.893	.038	.948	.869	.878	.027	.925	.863	.879	.040	.934	.865	.892	.040
Ours (*RGB&D*)	**.926**	**.885**	**.906**	.038	**.959**	**.892**	**.905**	**.022**	**.929**	**.876**	**.896**	**.039**	**.949**	**.908**	**.934**	**.030**

5 Conclusion

To tackle the new problem of weakly-supervised RGB-D salient object detection, we propose in this paper an end-to-end approach based on iterative updates of the internal pseudo-labels. This allows us to leverage depth information in eliminating non-salient background noises and generating reliable depth-refined pseudo-labels. Textual semantics is incorporated in the fill-in-the-blank fashion, which is used to estimate the confidence scores of pseudo-labels. Extensive experiments demonstrate the effectiveness and efficiency of our approach. In addition, our method is very generic and can be easily adapted to fully-supervised and unsupervised paradigms. In these scenarios, our variants also obtain superior performance over the existing state-of-the-art dedicated methods.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009.

[2] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015.

[3] Chenglizhao Chen, Jipeng Wei, Chong Peng, and Hong Qin. Depth-quality-aware salient object detection. *IEEE Transactions on Image Processing*, 30:2350–2363, 2021.

[4] Chenglizhao Chen, Jipeng Wei, Chong Peng, Weizhong Zhang, and Hong Qin. Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion. *IEEE Transactions on Image Processing*, 29:4296–4307, 2020.

[5] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for RGB-D salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3051–3060, 2018.

[6] Hao Chen and Youfu Li. Three-stream attention-aware network for RGB-D salient object detection. *IEEE Transactions on Image Processing*, 28(6):2825–2835, 2019.

[7] Hao Chen, Youfu Li, and Dan Su. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. *Pattern Recognition*, 86:376–385, 2019.

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 801–818, 2018.

[9] Shuhan Chen and Yun Fu. Progressively guided alternate refinement network for RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision*, 2020.

[10] Zuyao Chen, Runmin Cong, Qianqian Xu, and Qingming Huang. DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection. *IEEE Transactions on Image Processing*, 2020.

[11] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *Proceedings of International Conference on Internet Multimedia Computing and Service*, pages 23–27, 2014.

[12] Runmin Cong, Jianjun Lei, Huazhu Fu, Junhui Hou, Qingming Huang, and Sam Kwong. Going from RGB to RGBD saliency: A depth-guided transformation model. *IEEE Transactions on Systems, Man, and Cybernetics*, 50(8):3627–3639, 2020.

[13] Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Nam Ling. HSCS: Hierarchical sparsity based co-saliency detection for RGBD images. *IEEE Transactions on Multimedia*, 21(7):1660–1671, 2018.

[14] Runmin Cong, Jianjun Lei, Changqing Zhang, Qingming Huang, Xiaochun Cao, and Chunping Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. *IEEE Signal Processing Letters*, 23(6):819–823, 2016.

[15] Karthik Desingh, K Madhava Krishna, Deepu Rajan, and CV Jawahar. Depth really matters: Improving visual salient region detection with depth. In *British Machine Vision Conference*, 2013.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2018.

[17] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4548–4557, 2017.

[18] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 698–704, 2018.

[19] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[20] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In *Proceedings of the European Conference on Computer Vision*, 2020.

[21] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3052–3062, 2020.

[22] Jingfan Guo, Tongwei Ren, and Jia Bei. Salient object detection for RGB-D image via saliency evolution. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2016.

[23] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. *IEEE Transactions on Cybernetics*, 48(11):3171–3183, 2017.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[25] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017.

[26] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised salient object detection by learning a classifier-driven map generator. *IEEE Transactions on Image Processing*, 28(11):5435–5449, 2019.

[27] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *IEEE International Conference on Image Processing*, pages 1115–1119, 2014.

[28] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[29] ByoungChul Ko, Soo Yeong Kwak, and Hyeran Byun. SVM-based salient region(s) extraction method for image retrieval. In *Proceedings of the 17th International Conference on Pattern Recognition*, volume 2, pages 977–980, 2004.

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[31] Chongyi Li, Runmin Cong, Sam Kwong, Junhui Hou, Huazhu Fu, Guopu Zhu, Dingwen Zhang, and Qingming Huang. ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection. *IEEE Transactions on Cybernetics*, PP(99):1–13, 2020.

[32] Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy. RGB-D salient object detection with cross-modality modulation and selection. In *Proceedings of the European Conference on Computer Vision*, 2020.

[33] Gongyang Li, Zhi Liu, Linwei Ye, Yang Wang, and Haibin Ling. Cross-modal weighting network for RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision*, 2020.

[34] Guanbin Li, Yuan Xie, and Liang Lin. Weakly supervised salient object detection using image labels. In *The AAAI Conference on Artificial Intelligence*, 2018.

[35] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2976–2983, 2013.

[36] Fangfang Liang, Lijuan Duan, Wei Ma, Yuanhua Qiao, Zhi Cai, and Laiyun Qing. Stereoscopic saliency model using contrast and depth-guided-background prior. *Neurocomputing*, 275:2227–2238, 2018.

[37] Guibiao Liao, Wei Gao, Qiuping Jiang, Ronggang Wang, and Ge Li. MMNet: Multi-stage and multi-scale fusion network for RGB-D salient object detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2436–2444, 2020.

[38] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for RGB-D saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13756–13765, 2020.

[39] Ao Luo, Xin Li, Fan Yang, Zhicheng Jiao, Hong Cheng, and Siwei Lyu. Cascade graph neural networks for RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision*, 2020.

[40] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2014.

[41] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. DeepUSPS: Deep robust unsupervised saliency prediction via self-supervision. In *Advances in Neural Information Processing Systems*, pages 204–214, 2019.

[42] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 454–461, 2012.

[43] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision*, 2020.

[44] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. RGBD salient object detection: a benchmark and algorithms. In *Proceedings of the European Conference on Computer Vision*, pages 92–109, 2014.

[45] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[46] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7254–7263, 2019.

[47] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9060–9069, 2020.

[48] Yao Qin, Huchuan Lu, Yiqun Xu, and He Wang. Saliency detection via cellular automata. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 110–119, 2015.

[49] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. RGBD salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5):2274–2285, 2017.

[50] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. Exploiting global priors for RGB-D saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 25–32, 2015.

[51] Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):769–779, 2013.

[52] Riku Shigematsu, David Feng, Shaodi You, and Nick Barnes. Learning RGB-D salient object detection using background enclosure, depth contrast, and top-down features. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2749–2757, 2017.

[53] Wei-Chih Tu, Shengfeng He, Qingxiong Yang, and Shao-Yi Chien. Real-time salient object detection with a minimum spanning tree. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2334–2342, 2016.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008, 2017.

[55] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017.

[56] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4894–4903, 2018.

[57] Siyue Yu, Bingfeng Zhang, Jimin Xiao, and Eng Gee Lim. Structure-consistent weakly supervised salient object detection with local saliency coherence. In *The AAAI Conference on Artificial Intelligence*, 2021.

[58] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, and Yizhou Yu. Multi-source weak supervision for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6074–6083, 2019.

[59] Dingwen Zhang, Junwei Han, Gong Cheng, and Ming-Hsuan Yang. Weakly supervised object localization and detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[60] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4048–4056, 2017.

[61] Dingwen Zhang, Junwei Han, Yu Zhang, and Dong Xu. Synthesizing supervision for learning deep saliency network without human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(7):1755–1769, 2019.

[62] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8582–8591, 2020.

[63] Jing Zhang, Jianwen Xie, and Nick Barnes. Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection. In *Proceedings of the European Conference on Computer Vision*, 2020.

[64] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12546–12555, 2020.

[65] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9029–9038, 2018.

[66] Miao Zhang, Sun Xiao Fei, Jie Liu, Shuang Xu, Yongri Piao, and Huchuan Lu. Asymmetric two-stream architecture for accurate RGB-D saliency detection. In *Proceedings of the European Conference on Computer Vision*, 2020.

[67] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for RGB-D saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3472–3481, 2020.

[68] Miao Zhang, Yu Zhang, Yongri Piao, Beiqi Hu, and Huchuan Lu. Feature reintegration over differential treatment: A top-down and adaptive fusion network for RGB-D salient object detection. In *ACM International Conference on Multimedia*, 2020.

[69] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for RGBD salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3927–3936, 2019.

[70] Jiawei Zhao, Yifan Zhao, Jia Li, and Xiaowu Chen. Is depth really necessary for salient object detection? In *ACM International Conference on Multimedia*, 2020.

[71] Wangbo Zhao, Jing Zhang, Long Li, Nick Barnes, Nian Liu, and Junwei Han. Weakly supervised video salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[72] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time RGB-D salient object detection. In *Proceedings of the European Conference on Computer Vision*, 2020.

[73] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

[74] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. RGB-D salient object detection: A survey. *Computational Visual Media*, pages 1–33, 2021.

[75] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. PDNet: Prior-model guided depth-enhanced network for salient object detection. In *IEEE International Conference on Multimedia and Expo*, pages 199–204, 2019.

[76] Chunbiao Zhu, Ge Li, Xiaoqiang Guo, Wenmin Wang, and Ronggang Wang. A multilayer backpropagation saliency detection algorithm based on depth mining. In *International Conference on Computer Analysis of Images and Patterns*, pages 14–23, 2017.

[77] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. An innovative salient object detection using center-dark channel prior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1509–1515, 2017.

[78] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2814–2821, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See supplementary material.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This work does not involve any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sec. 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Sec. 4.3.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] Our code and dataset will be publicly available upon paper acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Sec. 4.1, Sec. 4.2 and Sec. 4.3.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See supplementary material.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Sec. 4.2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Sec. 4.1
   (b) Did you mention the license of the assets? [Yes] The relevant assets used in this paper are publicly available.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]