Energy-Based Generative Cooperative Saliency Prediction

Jing Zhang[1], Jianwen Xie[2], Zilong Zheng[3], Nick Barnes[1]

[1] The Australian National University [2] Cognitive Computing Lab, Baidu Research [3] UCLA
zjnwpu@gmail.com, {jianwen, zilongzheng0318}@ucla.edu, nick.barnes@anu.edu.au

Abstract

Conventional saliency prediction models typically learn a deterministic mapping from images to the corresponding ground truth saliency maps. In this paper, we study the saliency prediction problem from the perspective of generative models by learning a conditional probability distribution over saliency maps given an image, and treating the saliency prediction as a sampling process. Specifically, we propose a generative cooperative saliency prediction framework, where a conditional latent variable model and a conditional energy-based model are jointly trained to predict saliency in a cooperative manner. The latent variable model serves as a fast but coarse predictor to efficiently produce an initial prediction, which is then refined by the iterative Langevin revision of the energy-based model that serves as a fine predictor. Such a coarse-to-fine cooperative saliency prediction strategy offers the best of both worlds. We further generalize our framework to weakly supervised saliency prediction with a cooperative learning while recovering strategy. Experimental results show that our generative model achieves both state-of-the-art performance and reliable sampling processing exploring.

1 Introduction

As a class-agnostic segmentation task, salient object detection has attracted a lot of attention in the computer vision community for its close relationship to human visual perception. A *salient region* is a visually distinctive scene region that can be located rapidly and with little human effort. Salient object detection is commonly treated as a pixel-wise binary output of a deterministic prediction model in most recent works (Wu, Su, and Huang 2019a; Qin et al. 2019; Wu, Su, and Huang 2019b; Wei, Wang, and Huang 2020; Wang et al. 2019; Xu et al. 2021).

Despite the success of those recent models, the "point estimation process" caused by the deterministic nature of the network has prevented them from fully exploring the "sampling process" of saliency generation. Saliency detection is subjective (Itti, Koch, and Niebur 1998) and will be affected by a set of factors, including biological (*e.g.*, contrast sensitivity), and contextual (*e.g.*, task, experience, interests), *etc.* In this way, it is more reasonable to model saliency as a sampling process, representing the individual preference of each observer towards the same scene (Itti, Koch, and Niebur 1998; Zhang et al. 2020a).

Generative models (Goodfellow et al. 2014; Kingma and Welling 2013) have demonstrated their ability to produce multiple plausible outputs given the same input (Zhu et al. 2017; Lee et al. 2020). In this work, we fit the saliency detection task into generative model based framework, where observed images are input conditions, and the goal of the framework is to generate a series of saliency maps, representing the "subjective nature" of saliency. Additionally, we propose to model the conditional distribution of saliency maps *explicitly* in an energy-based model (EBM), which is well-known for its flexibility and effectiveness in distribution paramterization (Xie et al. 2018; Nijkamp et al. 2019). For our saliency detection task we find that the energy-based framework can learn a reliable and generalizable cost function, which can refine the prediction from the latent variable model by pushing it to the local mode in the energy landscape of the energy-based model. It is worth noting that, although (Zhang et al. 2020a, 2021) used conditional variational auto-encoders (Kingma and Welling 2013; Sohn, Lee, and Yan 2015a) to model labeling variants for saliency detection, they directly find a local minima with the latent variable model. Going beyond that, in this paper, we further search in the neighbourhood of the local minima with an EBM to obtain a refined prediction.

The energy-based model (EBM) learns an energy function by maximum likelihood estimation (MLE) and generates outputs via an iterative sampling process using Markov Chain Monte Carlo (MCMC) given the updated energy function. However, this framework commonly suffers from difficulties in modeling high-dimensional data from the computational expensiveness of MCMC. Inspired by prior success of cooperative training in unconditional image generation (Xie et al. 2018), in this work, we bring the best of both worlds, namely saliency prediction models and EBMs, into one framework and propose a *conditional cooperative saliency prediction network* for modeling the saliency distribution. Specifically, the model consists of a conditional energy-based model (EBM) whose energy function is parameterized by a bottom-top neural network, and a conditional latent variable model (LVM) whose transformation function is parameterized by an encoder-decoder framework, where the prior high-performance prediction model comes in. By using a latent variable model as an ancestral sampler to approximate or initialize the MCMC computational process for efficient sampling, the EBM can be learned efficiently. The energy function in the EBM, in turn, can be used to refine the LVM's samples, achieving a coarse-to-fine generative saliency detection.

Moreover, based on the conditional cooperative network, we introduce a *cooperative learning while cooperative recovering* strategy for weakly supervised saliency learning, where each training image is associated with a partially observed annotation (*e.g.*, scribble (Zhang et al. 2020b)). At each learning iteration, the incomplete saliency ground truth is firstly recovered in the low-dimensional latent space of the latent variable model via inference, and then refined by pushing it to the local mode in the energy landscape of the energy-based model. Although LVM is used in (Zhang et al. 2020a, 2021), there exists no such extension in these models.

Our contributions can be summarized as: (i) We study generative modeling of saliency prediction, and formulate this problem using EBM and LVM respectively, which are new angles to model and solve saliency prediction. (ii) We propose a generative cooperative saliency prediction framework, which jointly trains the LVM predictor and the EBM predictor in a cooperative learning scheme to offer reliable and efficient saliency prediction. (iii) We generalize our method to weakly supervised saliency prediction scenarios with incomplete annotations by proposing the *cooperative learning while recovering* algorithm, where we train the model and simultaneously recover the unlabeled area. (iv) We provide strong empirical results in both fully and weakly supervised settings to corroborate our framework.

2 Related Work

Fully/Weakly Supervised Saliency Models. Existing fully supervised saliency prediction models (Wang et al. 2018; Liu, Han, and Yang 2018; Wei, Wang, and Huang 2020; Liu et al. 2019; Qin et al. 2019; Wu, Su, and Huang 2019b,a; Wang et al. 2019; Wang et al. 2019; Wei et al. 2020; Xu et al. 2021) mainly focus on exploring image context information and generating structure-preserving predictions. (Wu, Su, and Huang 2019b; Wang et al. 2019; Wang et al. 2019; Wang et al. 2018; Liu, Han, and Yang 2018; Liu et al. 2019; Wu, Su, and Huang 2019a; Xu et al. 2021) introduced saliency prediction models by effectively integrating higher- and lower-level features. (Wei, Wang, and Huang 2020; Wei et al. 2020) proposed an edge-aware loss term to penalize errors along object boundaries. (Zhang et al. 2020a) presented a conditional variational auto-encoders (Kingma and Welling 2013; Jimenez Rezende, Mohamed, and Wierstra 2014) based stochastic RGB-D saliency detection network. Similarly, we introduce a cooperative learning pipeline to achieve probabilistic coarse-to-fine RGB saliency detection, where the latent variable model produces coarse prediction, which is then refined by the energy-based model.

The weakly supervised saliency models (Wang et al. 2017; Li, Xie, and Lin 2018a; Nguyen et al. 2019; Zhang et al. 2020b) learn saliency from easy-to-obtain weak labels, including image-level labels (Wang et al. 2017; Li, Xie, and Lin 2018b), noisy labels (Nguyen et al. 2019; Zhang et al. 2018; Zhang, Han, and Zhang 2017) or partial scribble labels (Zhang et al. 2020b). Although a probabilistic model was explored in (Zhang, Xie, and Barnes 2020), they used a generative model for noise modeling where the latent variable is to model the noise distribution. In this paper, for the weakly supervised task, we use a latent variable to model the distribution of the hidden clean saliency map.

Generative Cooperative Networks. Deep energy-based generative models (Xie et al. 2016), with energy functions parameterized by modern convolutional neural networks, are capable of modeling the probability density of high-dimensional data, which has been applied to image generation (Nijkamp et al. 2019), video generation (Xie, Zhu, and Wu 2019), and *etc.* The maximum likelihood learning of the energy-based model typically requires iterative MCMC sampling, which is computational challenging. To relieve the computational burden of MCMC, the Generative Cooperative Networks (CoopNets) in (Xie et al. 2018) propose to learn a separate latent variable model (*i.e.* a generator) to serve as an efficient approximate sampler for training the energy-based model. Our paper proposes a conditional model under the cooperative learning framework for visual saliency prediction. Further, we generalize to weakly supervised learning by proposing a *cooperative learning while recovering* algorithm. In this way, we can learn from incomplete data for weakly supervised saliency prediction.

Conditional Deep Generative Models. Our framework belongs to the family of conditional generative models, which also include conditional generative adversarial networks (CGANs) (Mirza and Osindero 2014) and conditional variational auto-encoders (CVAEs) (Sohn, Lee, and Yan 2015a). Different from existing CGAN-based conditional generative models (Luc et al. 2016; Zhang et al. 2018; Xue et al. 2017; Pan et al. 2017; Yu and Cai 2018; Hung et al. 2018; Souly, Spampinato, and Shah 2017), which use GANs to detect higher-order inconsistency between ground truth and the prediction, or CVAEs based models (Kohl et al. 2018; Zhang et al. 2020a, 2021) in which a latent variable model representing an implicit density is learned, our model learns an explicit density via energy-based modeling. More importantly, our model allows an additional refinement for the latent variable model during prediction, which is sorely lacking in both CGANs and CVAEs frameworks.

3 Cooperative Saliency Prediction

We will first present two types of generative modeling of saliecny prediction. Then, we propose a novel saliency prediction framework, in which the energy-based model and the latent variable model are jointly trained in a generative cooperative manner, such that they can help each other for better saliency prediction in terms of efficiency and accuracy.

3.1 EBM as a Fine but Slow Predictor

Let X be an image, and Y be its saliency map. The energy-based model $p_\theta(Y|X)$ defines a distribution of saliency Y

given an image X by:

$$p_\theta(Y|X) = \frac{p_\theta(Y,X)}{\int p_\theta(Y,X)dY} = \frac{1}{Z(X;\theta)}\exp[-U_\theta(Y,X)], \tag{1}$$

where the energy function $U_\theta(Y,X)$, parameterized by a bottom-up neural network, maps the input image-saliency pair to a scalar, and θ represents the network parameters. $Z(X;\theta) = \int \exp[-U_\theta(Y,X)]dY$ is the normalizing constant. When U_θ is learned and an image X is given, the prediction of saliency Y can be achieved by Langevin sampling (Neal 2012) $Y \sim p_\theta(Y|X)$, which makes use of the gradient of the energy function and iterates the following step:

$$Y_{\tau+1} = Y_\tau - \frac{\delta^2}{2}\frac{\partial U_\theta(Y_\tau,X)}{\partial Y} + \delta\Delta_\tau, \Delta_\tau \sim N(0,I_D), \tag{2}$$

where τ indexes the Langevin time steps, and δ is the step size. Langevin dynamics (Neal 2012) is equivalent to a stochastic gradient descent algorithm that seeks to find the minimum of the objective function defined by $U_\theta(Y,X)$. The Gaussian noise term Δ_τ is a Brownian motion that prevents gradient descent from being trapped by local minima of $U_\theta(Y,X)$. The prediction process via Langevin dynamics in Eq. (2) can be considered as finding Y to minimize the cost $U_\theta(Y,X)$ given input X. Such a framework can learn a reliable and generalizable cost function for saliency prediction. However, due to the iterative sampling process, EBM is a fine but slow saliency predictor.

3.2 LVM as a Coarse but Fast Predictor

Let h be a latent Gaussian noise vector, $G_\alpha(X,h)$ be a mapping function parameterized by a noise-injected encoder-decoder network with skip connections. α contains all the learning parameters in the network. The latent variable model is given by:

$$h \sim N(0,I_d), Y = G_\alpha(X,h) + \epsilon, \epsilon \sim N(0,\sigma^2 I_D), \tag{3}$$

which defines an implicit conditional distribution $p_\alpha(Y|X) = \int p_\alpha(Y|X,h)p(h)dh$ of saliency Y given an image X, where $p_\alpha(Y|X,h) = N(G_\alpha(X,h),\sigma^2 I_D)$. The saliency prediction can be achieved by an ancestral sampling by first sampling an injected Gaussian white noise vector h and then transforming it and the image X to a saliency map Y. Since the ancestral sampling is a direct mapping, it is faster than the iterative Langevin dynamics.

3.3 Cooperative Prediction with Two Predictors

We propose to predict image saliency by a cooperative sampling strategy, where we first use the coarse saliency predictor to generate an initial prediction $\hat{Y}$ via a non-iterative ancestral sampling, and then we use the fine saliency predictor to refine the initial prediction via iterative Langevin revision to obtain the revised saliency $\tilde{Y}$, *i.e.*,

$$\begin{aligned}&\hat{Y} = G_\alpha(X,\hat{h}), \hat{h} \sim N(0,I_d),\\ &\tilde{Y}_{\tau+1} = \tilde{Y}_\tau - \frac{\delta^2}{2}\frac{\partial U_\theta(\tilde{Y}_\tau,X)}{\partial \tilde{Y}} + \delta N(0,I_D), \tilde{Y}_0 = \hat{Y}.\end{aligned} \tag{4}$$

We call this cooperative sampling based coarse-to-fine prediction. In this way, we take both advantages of these two saliency predictors in the sense that the fine saliency predictor (*i.e.*, Langevin sampler) is initialized by the efficient coarse saliency predictor (*i.e.*., ancestral sampler), while the coarse saliency predictor is refined by the accurate fine saliency predictor that aims to minimize a cost function U_θ.

Since our conditional model represents a stochastic mapping, the prediction is stochastic as well. To evaluate the learned model on saliency prediction tasks, we can draw multiple h's from the prior $N(0,I_d)$ and use their average to generate $\hat{Y}$, then a Langevin dynamics with noise disabled (*i.e.*, gradient descent) is performed to push $\hat{Y}$ to its nearest local minimum $\tilde{Y}$ based on the learned energy function. The resulting $\tilde{Y}$ is treated as a prediction of our model.

3.4 Cooperative Training of Two Predictors

We use the cooperative training method to learn the parameters of the two predictors. At each iteration, we first generate synthetic examples via the cooperative sampling strategy shown in Eq. (4), and then the synthetic examples are used to compute the gradients to update both predictors.

MCMC-based Maximum Likelihood Estimation (MLE) for the Fine Saliency Predictor. Given a training dataset $\{(X_i,Y_i)\}_{i=1}^n$, we train the fine saliency predictor via MLE, which maximizes the log-likelihood of the data $L(\theta) = \frac{1}{n}\sum_{i=1}^n \log p_\theta(Y_i|X_i)$, whose gradient is $\Delta\theta = \frac{1}{n}\sum_{i=1}^n\{E_{p_\theta(Y|X_i)}[\frac{\partial}{\partial\theta}U_\theta(Y,X_i)] - \frac{\partial}{\partial\theta}U_\theta(Y_i,X_i)\}$. We rely on the cooperative sampling in Eq. (4) to sample $\tilde{Y}_i \sim p_\theta(Y|X_i)$ to approximate the gradient:

$$\Delta\theta \approx \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}U_\theta(\tilde{Y}_i,X_i) - \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\theta}U_\theta(Y_i,X_i). \tag{5}$$

We can use Adam with $\Delta\theta$ to update θ. We denote $\Delta\theta(\{Y_i\},\{\tilde{Y}_i\})$ as a function of $\{Y_i\}$ and $\{\tilde{Y}_i\}$.

Maximum Likelihood Training of the Coarse Saliency Predictor by MCMC Teaching. Even though the fine saliency predictor learns from the training data, the coarse saliency predictor learns to catch up with the fine saliency predictor by treating $\{(X,\tilde{Y})\}_{i=1}^n$ as training examples. The learning objective is to maximize the log-likelihood of the samples drawn from $p_\theta(Y|X)$, *i.e.*, $L(\alpha) = \frac{1}{n}\sum_{i=1}^n \log p_\alpha(\tilde{Y}_i|X_i)$, whose gradient can be computed by

$$\Delta\alpha = \sum_{i=1}^n E_{h\sim p_\alpha(h|Y_i,X_i)}\left[\frac{\partial}{\partial\alpha}\log p_\alpha(Y_i,h|X_i)\right]. \tag{6}$$

This leads to an MCMC-based solution that iterates (i) an inference step: inferring latent $\tilde{h}$ by sampling from posterior distribution $h \sim p_\alpha(h|Y,X)$ via Langevin dynamics, which iterates the following:

$$h_{\tau+1} = h_\tau + \frac{\delta^2}{2}\frac{\partial}{\partial h}\log p_\alpha(Y,h_\tau|X) + \delta\Delta_\tau, \tag{7}$$

where $\Delta_\tau \sim N(0,I_d)$, $\frac{\partial}{\partial h}\log p_\alpha(Y,h_\tau|X) = \frac{1}{\sigma^2}(Y - G_\alpha(X,h_\tau))\frac{\partial}{\partial h}G_\alpha(X,h_\tau) - h_\tau$, and (ii) a learning step: with

$\{\tilde{h}_i, \tilde{Y}_i, X_i\}$, we update α via Adam optimizer with

$$\Delta\alpha \approx \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma^2}(\tilde{Y}_i - G_\alpha(X_i, \tilde{h}_i))\frac{\partial}{\partial\alpha}G_\alpha(X_i, \tilde{h}_i). \quad (8)$$

Since G_α is parameterized by a differentiable neural network, both $\frac{\partial}{\partial h}G_\alpha(X, h_\tau)$ in Eq. (7) and $\frac{\partial}{\partial\alpha}G_\alpha(X_i, \tilde{h}_i)$ in Eq. (8) can be efficiently computed by back-propagation.

4 Weakly Supervised Saliency Prediction

In Section 3, the framework is trained from fully-observed training data. In this section, we want to show that our generative framework can be modified to handle the scenario in which each image X_i only has a partial pixel-wise annotation Y_i', *e.g.*, scribble annotation (Zhang et al. 2020b). Since the saliency map for each training image is incomplete, directly applying the algorithm to the incomplete training data can lead to a failure of learning the distribution of saliency given an image. However, generative models are good at data recovery, therefore they can learn to recover the incomplete data. In our framework, we will leverage the recovery powers of both EBM and LVM to deal with the incomplete data in our cooperative learning algorithm, and this will lead to a novel weakly supervised saliency prediction framework.

To learn from incomplete data, our algorithm alternates the cooperative learning step and the cooperative recovery step. Both steps need a cooperation between EBM and LVM. The cooperative learning step is the same as the one used for fully observed data, except that it treats the recovered saliency maps, which are generated from the cooperative recovery step, as training data in each iteration. The following is the cooperative recovery step, which consists of two substeps driven by the LVM and EBM respectively:

(i) Recovery by LVM in Latent Space. Given an image X_i and its incomplete saliency map Y_i', the recovery of the missing part of Y_i' can be achieved by first inferring the latent variable h_i' based on partially observed saliency information via $h_i' \sim p_\alpha(h|Y_i', X_i)$, and then generating $\hat{Y}_i' = G_\alpha(X_i, h_i')$ with the inferred h_i'. Let O_i be a binary mask, with the same size as Y', indicating the locations of visible annotations in Y_i'. O_i varies for different Y_i' and can be extracted from Y_i'. The Langevin dynamics for recovery iterates the same step in Eq. (7) except that $\frac{\partial}{\partial h}\log p_\alpha(Y', h_\tau|X) = \frac{1}{\sigma^2}(O \circ (Y - G_\alpha(X, h_\tau)))\frac{\partial}{\partial h}G_\alpha(X, h_\tau) - h_\tau$, where $\circ$ denotes element-wise matrix multiplication operation.

(ii) Recovery by EBM in Energy Landscape. With the initial recovered result $\hat{Y}'$ from the coarse saliency predictor p_α, the fine saliency predictor p_θ can further refine the result by running finite steps of Langevin dynamics in Eq. (2) initialized from $\hat{Y}'$ and obtain $\tilde{Y}'$. The underlying principle is that the initial recovery $\hat{Y}'$ might be just around the local modes of the energy function. A few steps of Langevin dynamics (*i.e.*, stochastic gradient descent) of p_θ, starting from $\hat{Y}_i'$, will push $\hat{Y}_i'$ to its nearby low energy mode of its potential fully observed version Y_i.

Cooperative Learning and Recovering. At each iteration t, we perform the above cooperative recovery of the training saliency map $\{Y'\}_{i=1}^n$ via $p_{\theta^{(t)}}$ and $p_{\alpha^{(t)}}$, while learning $p_{\theta^{(t+1)}}$ and $p_{\alpha^{(t+1)}}$ from $\{X_i, \tilde{Y}_i'^{(t)}\}_{i=1}^n$, where $\tilde{Y}_i'^{(t)}$ is the recovered saliency map at iteration t. The parameter θ is still updated via Eq. (5) except that we replace Y_i by $\tilde{Y}_i'$. That is, at each iteration, we use the recovered $\tilde{Y}_i'$, instead of the original Y_i, along with $\tilde{Y}_i$ to compute the gradient of log-likelihood, which is denoted by $\Delta\theta(\{\tilde{Y}_i'\}, \{\tilde{Y}_i\})$.

5 Technical Details

Latent Variable Model: The latent variable model (LVM) $G_\alpha(X, h)$ (with ResNet50 (He et al. 2016) as backbone) maps image X and latent variable h (we expand h to same spatial size of X) to a coarse saliency map $\hat{Y}$. Specifically, we adopt the decoder from MiDaS depth estimation (Ranftl et al. 2020) for its simplicity, which gradually aggregates the higher level features with lower level features with residual connections. We introduce the latent variable h to the bottleneck of the LVM by concatenating it with the highest level features, and then feed it to a 3 × 3 convolutional layer to obtain feature map of the same size as the original highest level feature. Details of the latent variable model are introduced in the supplementary material. As shown in Eq. 8, parameters of the latent variable model are updated with the revised prediction $\tilde{Y}$ from the EBM as supervision. To avoid error propagation from the EBM to the LVM in the early stage of training (where $\tilde{Y}$ is not very accurate), we introduce ground truth Y to LVM and gradually decrease its contribution to updating the LVM inspired by the dynamic supervision for knowledge distillation (Hinton, Vinyals, and Dean 2015), which encourages the student model (LVM in our case) to learn from ground truth first, and then gradually let the prediction of the teacher model (EBM in our case) supervises the student model. Specifically, we define an extra loss for the LVM as: $\lambda\mathcal{L}^s(G_\alpha(X, \tilde{h}), Y)$, where λ is linearly annealed to 0 while training, and $\mathcal{L}^s$ is the sturcture-aware loss function in (Wei, Wang, and Huang 2020).

Energy Function: The energy function $U_\theta(Y, X)$ represents the energy of input pair. Let `cks1-n` denote a `k×k` Convolution-BatchNorm-RELU layer with `n` filters and stride `1`, `fc-n` a fully connected layer with `n` filters. The energy network is composed of: `c3s1-32, c4s2-64, c4s2-128, c4s2-256, c4s1-1, fc-100`.

Implementation Details: We trained our model using PyTorch with a maximum of 30 epochs. Each image is rescaled to 352 × 352. ResNet50 (He et al. 2016) is chosen as backbone of the latent variable model. Empirically, we set the dimension of the latent space as $h = 8$. The learning rates of the latent variable model and EBM function are initialized to 5e-5 and 1e-3 respectively. We used Adam optimizer with momentum 0.9 and decrease the learning rate 10% after 20 epochs. It took 20 hours of training with batch size 7 using a single NVIDIA GeForce RTX 2080Ti GPU.

6 Experiments

Datasets: We used the DUTS dataset (Wang et al. 2017) for training the fully supervised model, and scribble annotation S-DUTS (Zhang et al. 2020b) for training the weakly

Table 1: Performance comparison with benchmark saliency prediction models, where "BkB" indicates the backbone, and "R34" is ResNet34 backbone (He et al. 2016), "R50" is the ResNet50 backbone (He et al. 2016).

			DUTS				ECSSD				DUT				HKU-IS				PASCAL-S				SOD			
Method	Year	BkB	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
			Deep Fully Supervised Models																							
PoolNet	2019	R50	.887	.840	.910	.037	.919	.913	.938	.038	.831	.748	.848	.054	.919	.903	.945	.030	.865	.835	.896	.065	.820	.804	.834	.084
BASNet	2019	R34	.876	.823	.896	.048	.910	.913	.938	.040	.836	.767	.865	.057	.909	.903	.943	.032	.838	.818	.879	.076	.798	.792	.827	.094
SCRN	2019	R50	.885	.833	.900	.040	.920	.910	.933	.041	.837	.749	.847	.056	.916	.894	.935	.034	.869	.833	.892	.063	.817	.790	.829	.087
F3Net	2020	R50	.888	.852	.920	.035	.919	.921	.943	.036	.839	.766	.864	.053	.917	.910	.952	.028	.861	.835	.898	.062	.824	.814	.850	.077
ITSD	2020	R50	.885	.840	.913	.041	.919	.917	.941	.037	.840	.768	.865	.061	.917	.904	.947	.031	.860	.830	.894	.066	.836	.829	.867	.076
LDF	2020	R50	.892	.861	.925	.034	.919	.923	.943	.036	.839	.770	.865	.052	.920	.913	.953	.028	.842	.768	.863	.064	-	-	-	-
UCNet+	2021	R50	.888	.860	.927	.034	.921	.926	.947	.035	.839	.773	.869	.051	.921	.919	.957	**.026**	.851	.825	.886	.069	.828	.827	.856	.076
PAKRN	2021	R50	.900	.876	.935	.033	**.928**	.930	.951	.032	.853	.796	.888	.050	.923	**.919**	.955	.028	.858	.838	.896	.067	.833	.836	.866	.074
Our_F	2021	R50	**.902**	**.877**	**.936**	**.032**	**.928**	**.935**	**.955**	**.030**	**.857**	**.798**	**.889**	**.049**	**.927**	.917	**.960**	**.026**	**.873**	**.846**	**.909**	**.058**	**.854**	**.850**	**.885**	**.064**
		Weakly Supervised Models																								
SSAL	2020	R50	.803	.747	.865	.062	.863	.865	.908	.061	.785	.702	.835	.068	.865	.858	.923	.047	.798	.773	.854	.093	.750	.743	.801	.108
SCWS	2021	R50	.841	**.818**	.901	.049	.879	.894	.924	.051	.813	.751	.856	.060	.883	.892	.938	.038	.821	.815	.877	.078	.782	.791	.833	.090
Our_W	2021	R50	**.847**	.816	**.902**	**.048**	**.896**	**.896**	**.934**	**.045**	**.817**	**.762**	**.864**	**.058**	**.894**	**.893**	**.943**	**.037**	**.834**	**.823**	**.886**	**.073**	**.803**	**.793**	**.849**	**.082**

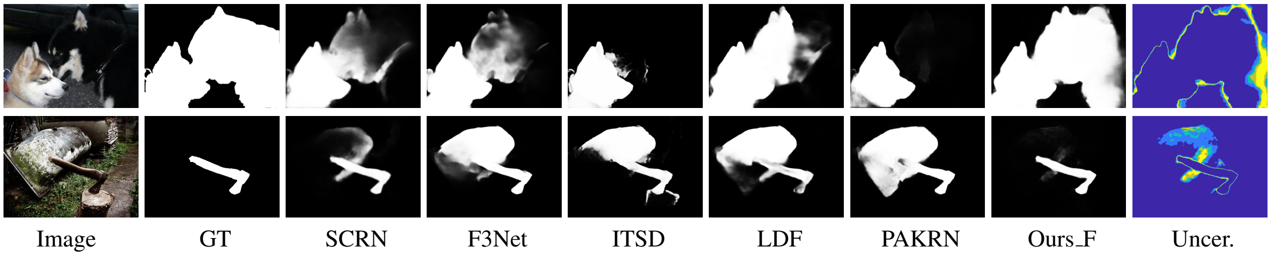

Figure 1: Visual comparison of fully supervised models.

supervised model. Testing images include 1) DUTS testing dataset, 2) ECSSD (Yan et al. 2013), 3) DUT (Yang et al. 2013), 4) HKU-IS (Li and Yu 2015), 5) PASCAL-S (Li et al. 2014) and 6) SOD dataset (Movahedi and Elder 2010).

Compared methods: We compared our method against state-of-the-art fully supervised saliency detection methods: PoolNet (Liu et al. 2019), BASNet (Qin et al. 2019), SCRN (Wu, Su, and Huang 2019b), F3Net (Wei, Wang, and Huang 2020), ITSD (Zhou et al. 2020), LDF (Wei et al. 2020), UCNet+ (Zhang et al. 2021) and PAKRN (Xu et al. 2021), where UCNet+ (Zhang et al. 2021) is the only generative model based framework. We also compare our weakly supervised solution with the scribble saliency detection models SSAL (Zhang et al. 2020b) and SCWS (Yu et al. 2021).

Evaluation Metrics: We evaluate performance of ours and compared methods with four saliency evaluation metrics, including: Mean Absolute Error ($\mathcal{M}$), mean F-measure (F_β), mean E-measure (E_ξ) (Fan et al. 2018) and S-measure (S_α) (Fan et al. 2017). Details about these metrics will be introduced in the supplementary material.

6.1 Comparison with Fully-supervised Models

Quantitative comparison: We evaluate performance of compared methods and ours and show results in Table 1, where "Ours_F" is the proposed fully supervised solution. We observe consistent performance improvement of "Ours_F" over six testing datasets compared with benchmark models, which clearly shows the advantage of our model. Note that, we use existing decoders (MiDaS decoder (Ranftl et al. 2020)) for the proposed latent variable model due to its easy implementation to focus on the learning pipeline. The ablation study on decoder (which will be discussed later) further explains our superior performance.

Qualitative comparison: As a generative model, we can model the sampling process of visual saliency, representing the subjective nature. In this way, two types of evaluation are necessary: 1) the deterministic prediction performance; and 2) the effectiveness of the stochastic predictions. For the former, we visualize predictions of ours and compared methods (SCRN (Wu, Su, and Huang 2019b), F3Net (Wei, Wang, and Huang 2020), ITSD (Zhou et al. 2020), LDF (Wei et al. 2020) and PAKRN (Xu et al. 2021)) in Fig. 1, where "Ours" is the EBM refined mean prediction from the LVM as discussed in Section 3.3. The visually better results of our model further validate our effectiveness. For the latter, we adopt uncertainty estimation as standard for stochastic prediction evaluation. Uncertainty (Kendall et al. 2017) is defined as ignorance of the model based on the provided training dataset. In our case, "uncertainty" is used to measure the subjective nature of saliency (Itti, Koch, and Niebur 1998; Zhang et al. 2021). In Fig. 1, "Uncer." is the uncertainty of our prediction, which is defined as the entropy (Kendall et al. 2017; Kendall and Gal 2017; Zhang et al. 2021, 2020a) of our model prediction "Ours". Note that, "blue" indicates confident predictions, and "yellow" represents less confident predictions. Fig. 1 shows that the deterministic one-to-one mapping models may lead to over-confident predictions, while the proposed method can also provide "Uncertainty" map to explain model confidence on it's prediction, which is beneficial in explaining model predictions.

Table 2: Performance comparison of ablation study related models.

	DUTS				ECSSD				DUT				HKU-IS				PASCAL-S				SOD			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
$G_\alpha(X)$	.878	.835	.918	.038	.916	.915	.946	.036	.826	.751	.862	.058	.912	.901	.952	.030	.856	.830	.899	.064	.829	.827	.871	.072
$G_\alpha(X,h)$	.897	.858	.932	.034	.918	.923	.946	.034	.837	.777	.882	.051	.914	.913	.957	.028	.863	.835	.900	.062	.831	.830	.874	.070
ITSD	.885	.840	.913	.041	.919	.917	.941	.037	.840	.768	.865	.061	.917	.904	.947	.031	.860	.830	.894	.066	.836	.829	.867	.076
ITSD_Ours	.914	.880	.945	.030	.938	.935	.959	.029	.860	.803	.901	.044	.933	.927	.971	.026	.875	.848	.921	.055	.845	.835	.880	.067
VGG16_Ours	.906	.876	.941	.032	.939	.933	.953	.030	.857	.799	.893	.048	.929	.923	.959	.027	.871	.844	.907	.058	.841	.838	.871	.066
Our_F	.902	.877	.936	.032	.928	.935	.955	.030	.857	.798	.889	.049	.927	.917	.960	.026	.873	.846	.909	.058	.854	.850	.885	.064

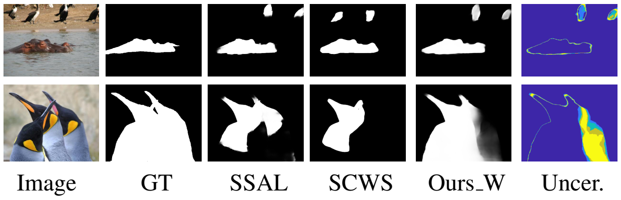

Figure 2: Visual comparison of the weakly supervised models.

Inference time and model size comparison: We have two main modules in our framework, namely a latent variable model and an energy-based model. The former takes the ResNet50 (He et al. 2016) backbone as encoder, and MiDaS (Ranftl et al. 2020) decoder for feature aggregation, leading to a model parameter size of 55M for the LVM. The latter adds 1M extra parameters to our cooperative learning framework. Our total model size is then 56M, which is comparable with mainstream saliency detection models, *e.g.*F3Net (Wei, Wang, and Huang 2020) has 48M parameters. Testing costs approximately 0.08 seconds for each iteration of sampling in the latent space, which is comparable to existing solutions as well. In this paper, we report mean predictions for performance evaluation, and we observe relative stable performance with each iteration of sampling for larger testing datasets, *e.g.* DUTS testing dataset (Wang et al. 2017), and slightly different performance for smaller testing datasets, *e.g.* SOD (Movahedi and Elder 2010) testing dataset.

6.2 Weakly Supervised Saliency Detection

We extend our solution to weakly supervised saliency detection with scribble annotation (Zhang et al. 2020b), and show model performance in Table 1, where "Our_W" in the "Weakly Supervised Models" section is our weak learning model. The better performance of our weak model again shows effectiveness of our solution. We further show the prediction from our model and existing solutions, namely SSAL (Zhang et al. 2020b) and SCWS (Yu et al. 2021) in Fig. 2. The accurate model prediction and reliable uncertainty map ("Uncer.") further validate our framework.

6.3 Ablation Study

We carried out the following experiments as shown in Table 2 to further analyse our solution.

Training the noise-free encoder-decoder G_α: We remove the latent variable h from our noise-injected encoder-decoder $G_\alpha(X,h)$ to obtain the deterministic noise-free encoder-decoder $G_\alpha(X)$. The result is shown as $G_\alpha(X)$. We observe inferior performance of $G_\alpha(X)$ compared with benchmark fully supervised saliency detection models, *i.e.*PAKRN (Xu et al. 2021). This is mainly because the decoder (Ranftl et al. 2020) we adopted is for depth estimation, and we use it in our latent variable model for its simple implementation. However, the consistently better performance of "Ours_F" with the cooperative learning pipeline (which is built upon $G_\alpha(X)$) compared with existing techniques validates the effectiveness of our solution.

Training the noise-injected encoder-decoder $G_\alpha(X,h)$: We treat the stochastic saliency encoder-decoder $G_\alpha(X,h)$ as our final model, where the latent variable h is updated through Langevin dynamics as shown in Eq. 7. In this way, our method performs alternating back-propagation (Han et al. 2017). The result is shown as $G_\alpha(X,h)$. Compared with $G_\alpha(X)$, $G_\alpha(X,h)$ achieves better performance, which illustrates the effectiveness of the latent variable model.

Decoder and encoder ablation: We replace our decoder with decoder from existing SOD model, ITSD (Zhou et al. 2020) in particular[1], and show its performance as "ITSD_Ours". Further, to ablate the contribution of the backbone network, we replace the ResNet50 (He et al. 2016) backbone with VGG16 (Simonyan and Zisserman 2014) and show model performance as "VGG16_Ours". The consistently better performance of "ITSD_Ours" compared with the original "ITSD" explains our superior performance. The relatively stable model performance with different backbones, *e.g.*VGG16 (Simonyan and Zisserman 2014) and ResNet50 (He et al. 2016) validates the robustness of our model with respect to backbone networks.

6.4 Alternative Uncertainty Estimation Methods

To model the subjective nature of saliency, or uncertainty of saliency, instead of designing a deterministic one-to-one mapping framework, we aim to estimate the conditional distribution of saliency maps given input image X. To achieve this, generative models (Goodfellow et al. 2014; Kingma and Welling 2013) are straightforward solutions. We design two alternative generator network based saliency detection pipelines with CVAEs (Sohn, Lee, and Yan 2015b) and CGANs (Mirza and Osindero 2014) and performance is shown as "CVAE" and "CGAN" in Table 3 respectively. For the "CVAE" model, we use the same learning pipeline as (Zhang et al. 2021), except that we replace it's decoder with our decoder (Ranftl et al. 2020).

For "CGAN", we optimize the min-max game of conventional generative adversarial network with generator G

[1] We choose ITSD (Zhou et al. 2020) for decoder ablation due to its easily implemented code and state-of-the-art performance.

Table 3: Performance comparison with benchmark saliency prediction models.

	DUTS				ECSSD				DUT				HKU-IS				PASCAL-S				SOD			
Method	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
CVAE	.890	.849	.925	.036	.919	.918	.948	.034	.836	.761	.868	.056	.918	.906	.955	.028	.863	.835	.902	.062	.838	.830	.878	.071
CGAN	.888	.849	.927	.035	.917	.914	.944	.036	.837	.764	.871	.054	.917	.908	.955	.028	.865	.839	.906	.059	.836	.833	.874	.072
MCD	.881	.842	.918	.038	.917	.917	.944	.036	.828	.753	.859	.057	.915	.908	.951	.030	.863	.837	.902	.062	.834	.831	.868	.073
ENS	.885	.841	.921	.037	.921	.917	.948	.035	.831	.752	.862	.057	.916	.901	.952	.030	.858	.827	.897	.065	.835	.828	.872	.073
Our_F	**.902**	**.877**	**.936**	**.032**	**.928**	**.935**	**.955**	**.030**	**.857**	**.798**	**.889**	**.049**	**.927**	**.917**	**.960**	**.026**	**.873**	**.846**	**.909**	**.058**	**.854**	**.850**	**.885**	**.064**

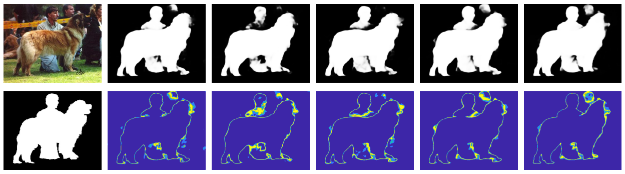

Figure 3: Predictions of alternative uncertainty estimation methods. 1^{st} row shows the input image and the mean predictions of each alternative models and ours, and the 2^{nd} row are ground truth and the corresponding uncertainty maps. From left to right are: Image/ground truth, "CVAE", "CGAN", "MCD", "ENS" and ours.

and discriminator D. Specifically, we use the same latent variable model $G_\alpha(X, h)$ as the generator of "CGAN". For the discriminator, we design a fully convolutional discriminator as (Hung et al. 2018) to distinguish the per-pixel input as real (ground truth) or fake (prediction). To train "CGAN", the discriminator is updated with discriminator loss $\mathcal{L}_{ce}(D(Y), \mathbf{1}) + \mathcal{L}_{ce}(D(G_\alpha(X, h)), \mathbf{0})$, where $\mathcal{L}_{ce}$ is the binary cross-entropy loss, $\mathbf{1}$ and $\mathbf{0}$ are all-one or all-zero feature maps of same spatial size as Y. The generator is updated with $\mathcal{L}_s(G_\alpha(X, h), Y) + \lambda_d \mathcal{L}_{ce}(D(G_\alpha(X, h)), \mathbf{1})$, where $\mathcal{L}_s$ is the structure-aware loss as in (Wei, Wang, and Huang 2020), and $\lambda_d \mathcal{L}_{ce}(D(G_\alpha(X, h))$ is the adversarial loss, and empirically we set $\lambda_d = 0.1$ (Hung et al. 2018).

We also design two ensemble based saliency detection network with Monte Carlo dropout (Gal and Ghahramani 2016) as approximate Bayesian posterior inference, and deep ensemble (Lakshminarayanan, Pritzel, and Blundell 2017) to produce multiple predictions with a multi-head decoder, and show their performance as "MCD" and "ENS" in Table 3 respectively. For "MCD", we add dropout to each level feature of the encoder within the noise-free encoder-decoder $G_\alpha(X)$ framework with dropout rate 0.3, and use dropout in both the training and testing processes. For "ENS", we attach five MiDaS decoder (Ranftl et al. 2020) to $G_\alpha(X)$, which are initialized differently, leading to five predictions in the end. For both the two types of ensemble based framework, same as our generative models, we define their mean predictions with $T = 10$ iterations of sampling during testing as their deterministic predictions, and entropy of the mean prediction is defined as the predictive uncertainty following (Detlefsen, Jørgensen, and Hauberg 2019).

Model Comparison: We show performance of alternative uncertainty estimation models in Table 3, and visualize the mean prediction and predictive uncertainty of each method in Fig. 3. For the CVAE based framework, designing the approximate inference network takes effort, and the imbalanced inference model may lead to the posterior collapse problem as discussed in (He et al. 2019), where the latent variable is independent of the prediction. For the CGAN based model, according to our experience, training is sensitive to the contribution of the adversarial loss. Further, it cannot infer the latent variable h, which makes the model not explanatory. For the ensemble (Lakshminarayanan, Pritzel, and Blundell 2017) or MC dropout (Gal and Ghahramani 2016) based solutions, they can hardly improve model performance, although the produced predictive uncertainty can explain model prediction to some extent. Compared those alternative methods, the proposed solution is stable, and we can infer the latent variable h without an extra encoder. Further, as we directly sample from the "truth" posterior distribution via Langevin dynamics based MCMC instead of the approximated distribution, we have no posterior collapse problem, leading to more reliable predictive uncertainty maps compared with alternative solutions.

6.5 Discussion

The advantage of combining LVM with EBM: Firstly, compared to other generative methods (*e.g.* VAE (Kingma and Welling 2013) or GAN (Goodfellow et al. 2014)), our EBM serves as a refinement engine. The iterative MCMC process refines the output provided by the LVM by minimizing the energy function, which is also to mimic reinforcement learning, where our LVM is a policy (non-iterative) and the energy function in the EBM is a cost (or negative value) function. Secondly, different from other generative methods that only keep the LVM for prediction, our system still has an EBM to further revise the prediction of the LVM, which has been proven effective with our extensive experiments.

7 Conclusion

We propose a generative saliency prediction model based on the conditional generative cooperative network, where a latent variable model and an energy-based model are jointly trained in a cooperative learning scheme to achieve coarse-to-fine saliency prediction. Moreover, we introduce a cooperative learning while recovering strategy and extend our model to weakly supervised saliency detection. Extensive results illustrate that our method can lead to both accurate deterministic predictions and reliable uncertainty maps representing model ignorance about its predictions. One drawback of existing sampling based uncertainty estimation methods (including ours) is the less effective structure-correlation modeling for the uncertainty map generation, where uncertainty of each pixel is treated independently. A more effective generative model with structure-correlated uncertainty representation is our next direction.

References

Detlefsen, N. S.; Jørgensen, M.; and Hauberg, S. 2019. Reliable training and estimation of variance networks. *arXiv preprint arXiv:1906.03260*.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A New Way to Evaluate Foreground Maps. In *Proc. IEEE Int. Conf. Comp. Vis.*, 4548–4557.

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 *Proc. IEEE Int. Joint Conf. Artificial Intell.*, 698–704.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proc. Int. Conf. Mach. Learn.*, 1050–1059.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Proc. Adv. Neural Inf. Process. Syst.*, 2672–2680.

Han, T.; Lu, Y.; Zhu, S.; and Wu, Y. 2017. Alternating Back-Propagation for Generator Network. In *Proc. AAAI Conf. Artificial Intelligence*.

He, J.; Spokoyny, D.; Neubig, G.; and Berg-Kirkpatrick, T. 2019. Lagging Inference Networks and Posterior Collapse in Variational Autoencoders. In *Proc. Int. Conf. Learning Representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 770–778.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. In *Proc. Adv. Neural Inf. Process. Syst. Workshop*.

Hung, W.-C.; Tsai, Y.-H.; Liou, Y.-T.; Lin, Y.-Y.; and Yang, M.-H. 2018. Adversarial Learning for Semi-supervised Semantic Segmentation. In *Proc. Brit. Mach. Vis. Conf.*

Itti, L.; Koch, C.; and Niebur, E. 1998. A Model of Saliency-based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20: 1254 – 1259.

Jimenez Rezende, D.; Mohamed, S.; and Wierstra, D. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proc. Int. Conf. Mach. Learn.*

Kendall, A.; Badrinarayanan, V.; ; and Cipolla, R. 2017. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. In *Proc. Brit. Mach. Vis. Conf.*

Kendall, A.; and Gal, Y. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Proc. Adv. Neural Inf. Process. Syst.*

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. In *Proc. Int. Conf. Learning Representations*.

Kohl, S. A.; Romera-Paredes, B.; Meyer, C.; De Fauw, J.; Ledsam, J. R.; Maier-Hein, K. H.; Eslami, S.; Rezende, D. J.; and Ronneberger, O. 2018. A Probabilistic U-Net for Segmentation of Ambiguous Images. *Proc. Adv. Neural Inf. Process. Syst.*

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Proc. Adv. Neural Inf. Process. Syst.* Curran Associates, Inc.

Lee, H.-Y.; Tseng, H.-Y.; Mao, Q.; Huang, J.-B.; Lu, Y.-D.; Singh, M.; and Yang, M.-H. 2020. Drit++: Diverse image-to-image translation via disentangled representations. *Int. J. Comp. Vis.*, 128(10): 2402–2417.

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

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

Li, G.; and Yu, Y. 2015. Visual saliency based on multiscale deep features. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 5455–5463.

Li, Y.; Hou, X.; Koch, C.; Rehg, J. M.; and Yuille, A. L. 2014. The Secrets of Salient Object Segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 280–287.

Liu, J.-J.; Hou, Q.; Cheng, M.-M.; Feng, J.; and Jiang, J. 2019. A Simple Pooling-Based Design for Real-Time Salient Object Detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Liu, N.; Han, J.; and Yang, M.-H. 2018. PiCANet: Learning Pixel-wise Contextual Attention for Saliency Detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 3089–3098.

Luc, P.; Couprie, C.; Chintala, S.; and Verbeek, J. 2016. Semantic Segmentation using Adversarial Networks. In *NIPS Workshop on Adversarial Training*.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. *CoRR*, abs/1411.1784.

Movahedi, V.; and Elder, J. H. 2010. Design and perceptual validation of performance measures for salient object segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. Workshops*, 49–56.

Neal, R. 2012. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*.

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 *Proc. Adv. Neural Inf. Process. Syst.*

Nijkamp, E.; Hill, M.; Zhu, S.-C.; and Wu, Y. N. 2019. Learning Non-Convergent Non-Persistent Short-Run MCMC Toward Energy-Based Model. *Proc. Adv. Neural Inf. Process. Syst.*

Pan, J.; Canton, C.; McGuinness, K.; O'Connor, N. E.; Torres, J.; Sayrol, E.; and Giro-i Nieto, X. a. 2017. SalGAN: Visual Saliency Prediction with Generative Adversarial Networks. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. Workshops*.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. BASNet: Boundary-Aware Salient Object Detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 7479–7489.

Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; and Koltun, V. 2020. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.*

Simonyan, K.; and Zisserman, A. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. Int. Conf. Learning Representations*.

Sohn, K.; Lee, H.; and Yan, X. 2015a. Learning Structured Output Representation using Deep Conditional Generative Models. In *Proc. Adv. Neural Inf. Process. Syst.*, 3483–3491.

Sohn, K.; Lee, H.; and Yan, X. 2015b. Learning Structured Output Representation using Deep Conditional Generative Models. In *Proc. Adv. Neural Inf. Process. Syst.*, 3483–3491.

Souly, N.; Spampinato, C.; and Shah, M. 2017. Semi Supervised Semantic Segmentation Using Generative Adversarial Network. In *Proc. IEEE Int. Conf. Comp. Vis.*, 5689–5697.

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. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 136–145.

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. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 3127–3135.

Wang, W.; Shen, J.; Cheng, M.-M.; and Shao, L. 2019. An Iterative and Cooperative Top-Down and Bottom-Up Inference Network for Salient Object Detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Wang, W.; Zhao, S.; Shen, J.; Hoi, S. C. H.; and Borji, A. 2019. Salient Object Detection With Pyramid Attention and Salient Edges. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 1448–1457.

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

Wei, J.; Wang, S.; Wu, Z.; Su, C.; Huang, Q.; and Tian, Q. 2020. Label Decoupling Framework for Salient Object Detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 13025–13034.

Wu, Z.; Su, L.; and Huang, Q. 2019a. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 3907–3916.

Wu, Z.; Su, L.; and Huang, Q. 2019b. Stacked Cross Refinement Network for Edge-Aware Salient Object Detection. In *Proc. IEEE Int. Conf. Comp. Vis.*

Xie, J.; Lu, Y.; Gao, R.; Zhu, S.-C.; and Wu, Y. N. 2018. Cooperative Training of Descriptor and Generator Networks. *IEEE Trans. Pattern Anal. Mach. Intell.*

Xie, J.; Lu, Y.; Zhu, S.-C.; and Wu, Y. 2016. A Theory of Generative ConvNet. In *Proc. Int. Conf. Mach. Learn.*, volume 48, 2635–2644.

Xie, J.; Zhu, S.-C.; and Wu, Y. N. 2019. Learning energy-based spatial-temporal generative convnets for dynamic patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*

Xu, B.; Liang, H.; Liang, R.; and Chen, P. 2021. Locate Globally, Segment Locally: A Progressive Architecture With Knowledge Review Network for Salient Object Detection. In *Proc. AAAI Conf. Artificial Intelligence*, 3004–3012.

Xue, Y.; Xu, T.; Zhang, H.; Long, R.; and Huang, X. 2017. SegAN: Adversarial Network with Multi-scale L_1 Loss for Medical Image Segmentation. *Neuroinformatics*, 16.

Yan, Q.; Xu, L.; Shi, J.; and Jia, J. 2013. Hierarchical saliency detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 1155–1162.

Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2013. Saliency Detection via Graph-Based Manifold Ranking. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 3166–3173.

Yu, H.; and Cai, X. 2018. Saliency detection by conditional generative adversarial network. In *Ninth International Conference on Graphic and Image Processing*, 253.

Yu, S.; Zhang, B.; Xiao, J.; and Lim, E. G. 2021. Structure-Consistent Weakly Supervised Salient Object Detection with Local Saliency Coherence. In *Proc. AAAI Conf. Artificial Intelligence*.

Zhang, D.; Han, J.; and Zhang, Y. 2017. Supervision by Fusion: Towards Unsupervised Learning of Deep Salient Object Detector. In *Proc. IEEE Int. Conf. Comp. Vis.*, 4068–4076.

Zhang, J.; Fan, D.-P.; Dai, Y.; Anwar, S.; Saleh, F.; Aliakbarian, S.; and Barnes, N. 2021. Uncertainty Inspired RGB-D Saliency Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*

Zhang, J.; Fan, D.-P.; Dai, Y.; Anwar, S.; Saleh, F. S.; Zhang, T.; and Barnes, N. 2020a. UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

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

Zhang, J.; Yu, X.; Li, A.; Song, P.; Liu, B.; and Dai, Y. 2020b. Weakly-Supervised Salient Object Detection via Scribble Annotations. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Zhang, J.; Zhang, T.; Dai, Y.; Harandi, M.; and Hartley, R. 2018. Deep Unsupervised Saliency Detection: A Multiple Noisy Labeling Perspective. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 9029–9038.

Zhang, X.; Zhu, X.; Zhang, . X.; Zhang, N.; Li, P.; and Wang, L. 2018. SegGAN: Semantic Segmentation with Generative Adversarial Network. In *2018 IEEE Fourth International Conference on Multimedia Big Data (BigMM)*, 1–5.

Zhou, H.; Xie, X.; Lai, J.-H.; Chen, Z.; and Yang, L. 2020. Interactive Two-Stream Decoder for Accurate and Fast Saliency Detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*

Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017. Toward multimodal image-to-image translation. *arXiv preprint arXiv:1711.11586*.