An Energy-Based Prior for Generative Saliency

Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li

Abstract—We propose a novel generative saliency prediction framework that adopts an informative energy-based model as a prior distribution. The energy-based prior model is defined on the latent space of a saliency generator network that generates the saliency map based on a continuous latent variables and an observed image. Both the parameters of saliency generator and the energy-based prior are jointly trained via Markov chain Monte Carlo-based maximum likelihood estimation, in which the sampling from the intractable posterior and prior distributions of the latent variables are performed by Langevin dynamics. With the generative saliency model, we can obtain a pixel-wise uncertainty map from an image, indicating model confidence in the saliency prediction. Different from existing generative models, which define the prior distribution of the latent variables as a simple isotropic Gaussian distribution, our model uses an energy-based informative prior which can be more expressive in capturing the latent space of the data. With the informative energy-based prior, we extend the Gaussian distribution assumption of generative models to achieve a more representative distribution of the latent space, leading to more reliable uncertainty estimation. We apply the proposed frameworks to both RGB and RGB-D salient object detection tasks with both transformer and convolutional neural network backbones. We further propose an adversarial learning algorithm and a variational inference algorithm as alternatives to train the proposed generative framework. Experimental results show that our generative saliency model with an energy-based prior can achieve not only accurate saliency predictions but also reliable uncertainty maps that are consistent with human perception.

Index Terms—Adversarial learning, energy-based models, generative models, Langevin dynamics, salient object detection, variational inference, vision transformer.

I. INTRODUCTION

SALIENT object detection [2], [3], [4], [5], [6], [7] (SOD), also known as visual saliency prediction, aims to highlight objects in images that attract more attention than their surrounding areas. With the emergence of deep convolutional neural networks, significant performance improvements have been achieved in the field of SOD [2], [4], [8], [9]. Traditional SOD models aim to learn a deterministic one-to-one mapping function from the image domain to the saliency domain using a set of training images along with their saliency annotations. These approaches often lead to point estimation techniques. The one-to-one deterministic mapping mechanism employed in the current SOD frameworks makes them unsuitable for exploring the pixel-wise uncertainty in predicting salient objects [10]. Uncertainty measures the level of confidence the model has in its predictions. Without an uncertainty indicator, a one-to-one deterministic mapping model is prone to being over-confident. Further, the saliency output of an image is subjective and heavily relies on human preference [9]. Therefore, compared with the deterministic model, a stochastic generative model is a more natural approach to model the stochastic nature of labeling by individuals and represent image saliency.

To account for the subjective nature of saliency, we propose a novel generative model for salient object detection. This model incorporates latent variables to capture randomness and uncertainty in the mapping from the image domain to the saliency domain. By doing so, it allows the model to generate stochastic saliency predictions, which can be useful for uncertainty estimation. There are mainly two types of generative models that have been widely used, namely the variational auto-encoder (VAE) [11] and the generative adversarial net (GAN) [12], which correspond to two different generative learning strategies to train latent variable models. To train a top-down latent variable generator, VAE [11] introduces an extra encoder to approximate the intractable posterior distribution of the latent variables, and trains the generator via a perturbation of maximum likelihood. Based on the VAE, [9] introduces a conditional variational autoencoder (CVAE) [13] for RGB-D salient object detection. However, the potential issue of posterior collapse [14], which is inherent within the VAE framework, can make stochastic predictions less effective in generating reliable uncertainty estimates. In contrast, GAN introduces a discriminator to distinguish the generated examples from the real examples, and trains the generator to fool the discriminator without an inference step. However, due to the absence of an explicit inference process in GANs, it can be difficult to handle inverse problems and may lead to the potential issue of mode collapse.

To address the potential issues like posterior collapse in VAEs and mode collapse in GANs, [16], [17] present the third learning strategy called alternating back-propagation (ABP). This method trains the generator with an inference process, which directly samples latent variables from the true posterior distribution using gradient-based Markov chain Monte Carlo (MCMC) [18] techniques such as Langevin dynamics [19], [20], [21]. It is worth noting that all the three generative models (i.e.,
The main goal of the existing (EABP), for salient object and ABP (row) of the conventional isotropic Gaussian prior based generative models (GAN [15], VAE [11] and ABP [16]) and the proposed EBM prior based models (EGAN, EVAE and EABP).

We aim to investigate generative SOD with an expressive and meaningful latent representation. To achieve this, instead of defining the prior distribution of the latent variables as an isotropic Gaussian distribution as in the conventional generative models [11], [15], [16], we assume that the latent variables follow an informative trainable energy-based prior distribution [22], [23], which is parameterized by a deep neural network. Such informative latent space modeling strategy with an energy-based model (EBM) as a prior can be plugged into existing latent variable models to achieve reliable latent space modeling, leading to EBM based GAN [12] (EGAN), EBM based VAE [11] (EVAE) and EBM based ABP [16] (EABP). For salient object detection, the informative latent space is the necessity for reliable saliency subjective nature modeling.

For EGAN, we introduce an extra inference model, where the prior of the latent variable is modeled with the EBM, and the posterior is updated via Langevin dynamics based MCMC [19]. For EVAE, the prior and posterior of the latent variable within the VAE framework [11] serve as the start point of our EBM prior and posterior models respectively, leading to an image-dependent warm-start of both models. The prior and posterior of EABP is the same as for EGAN. Differently, there exists extra discriminator in EGAN to achieve adversarial learning. Note that the generator of the three models can be trained by maximum likelihood estimation (MLE). The MLE algorithm relies on MCMC sampling to evaluate the intractable prior and posterior distributions of the latent variables. In this way, we do not suffer the posterior collapse issue as does the original VAE model [11], and the latent variable is more meaningful which is not an isotropic Gaussian but modeled with a deep neural network, that can be any distribution. We apply the proposed generative models to RGB and RGB-D salient object detection, and show that the generative frameworks with the EBM prior are powerful in representing the conditional distribution of object saliency given an image, leading to more reasonable uncertainty estimation [10] compared with the isotropic Gaussian distribution based counterparts (see Fig. 1).

This paper is an extended version of our conference paper [1]. In particular, we extend [1] in the following ways: (i) after we first introduce “EABP” from our original version, we then extend our EBM prior to GAN [15] and VAE [11], demonstrating that the proposed approach can be effective for a broad range of generative techniques; (ii) we implement the three types of EBM prior based generative models with both transformer (which is our conference version in [1]) and convolutional neural network (CNN) backbones (which is our extension) to verify the superiority of the proposed EBM prior based generative models for uncertainty estimation; (iii) we also apply the proposed EBM prior based generative models to existing saliency detection models [24] to explain the generalization ability of the proposed strategies; (iv) we extensively analyse the transformer backbone and CNN backbones w.r.t. both deterministic performance and uncertainty modeling, observing that although significant deterministic performance gaps exists for the two types of backbones, our EBM prior based generative models within both scenarios can generate reliable uncertainty maps to explain the corresponding predictions.

**II. RELATED WORK**

*Salient Object Detection:* The main goal of the existing deep fully-supervised RGB image-based salient object detection models [2], [3], [8], [25], [26], [27], [28], [29], [30] is to achieve structure preserving saliency prediction, by either sophisticated feature aggregation [25], [27], auxiliary edge detection [3], [8], [31], or resorting to structure-aware loss functions [2], [26]. [8] introduces an extra edge module and uses the interaction between the edge module and the detection module to optimize the two tasks at the same time. [2] adaptively selects
complementary features while performing multi-scale feature aggregation with a new structure-aware loss function for edge-aware saliency prediction. Similarly, [31] uses multiple supervisions and high-quality boundaries to guide the network to gradually optimize saliency prediction. [32] integrates the information of adjacent layers, as well as multi-scale information to retain the internal consistency of each category, i.e., salient foreground and non-salient background. With extra depth information, RGB-D salient object detection models [4], [6], [9], [33], [34], [35], [36], [37], [38], [39], [40], [41] mainly focus on effective multi-modal modeling, where both the appearance information from the RGB image and geometric information from depth are fused at various levels, for example, input level (early fusion models), output level (late fusion models), and feature level (cross-fusion models). Our paper solves the same problems, i.e., RGB and RGB-D salient object detection, by developing a new generative transformer-based saliency detection framework.

Vision Transformers: Transformer [42] is a set-to-set method based on a self-attention mechanism, which has achieved great success in natural language processing (NLP), and the breakthroughs in the NLP domain have sparked the interest of the computer vision community in developing vision transformers for different computer vision tasks, such as image classification [43], [44], object detection [44], [45], [46], [47], image segmentation [44], [46], [48], [49], object tracking [50], [51], pose estimation [52], [53], etc. Among vision transformers, DPT [49] adopts a U-shaped structure and uses ViT [43] as an encoder to perform semantic segmentation and monocular depth estimation. Swin [44] presents a hierarchical transformer with a shifted windowing scheme to achieve an efficient transformer network with high resolution images as input. Different from the above vision transformers that mainly focus on discriminative modeling and learning, our paper emphasizes generative modeling and learning of the vision transformer by involving latent variables and MCMC inference.

Dense Prediction with Generative Models: The goal of dense generative models is to produce both deterministic accurate model prediction and a meaningful stochastic uncertainty map representing model awareness of its prediction. Two types of generative models have been widely studied, namely VAEs [11], [13] and GANs [12]. VAEs use an extra encoder to constrain the distribution of the latent variable, and GANs design a discriminator to distinguish real samples and the generated samples. VAEs have been successfully applied to image segmentation [54], [55]. For saliency prediction, [56] adopts a VAE to model the image background, and separates salient objects from the background through reconstruction residuals. [9], [57] design CVAEs to model the subjective nature of saliency. GAN-based methods can be classified into two categories, depending on whether they are designed for fully-supervised or semi-supervised settings. The fully-supervised models [58], [59] use a discriminator to distinguish model predictions from the ground truth, while the semi-supervised models [60], [61] use the GAN to explore the information contained in unlabeled data. [62] uses a cooperative learning framework [63], [64], [65] for generative saliency prediction. [57] trains a single top-down generator in the ABP framework for RGB-D saliency prediction. Our model generalizes [57] by replacing the simple Gaussian prior by a learnable EBM prior and adopting a vision transformer generator for SOD.

Energy-based Models: Recent works have shown strong performance of data space EBMs [66], [67] in modeling high-dimensional complex dependencies, such as images [68], [69], [70], [71], [72], [73], [74], [75], [76], videos [77], [78], 3D shapes [79], [80], unordered point clouds [81] and trajectories [82], and also the effectiveness of latent space EBMs [22] in improving model expressivity for text [23], image [22], [83], and trajectory [84] generation. Our paper also learns a latent space EBM as the prior model but builds the EBM on top of a transformer generator for image-conditioned saliency map prediction. This leads to three EBM prior based generative saliency detection models to relax the “Gaussian assumption” of conventional latent variable models [11], [15], [16], [17].

III. GENERATIVE SALIENCY DETECTION

A. The Formulation

Generative saliency detection can be formulated as a conditional generative learning problem. We begin by introducing some notation. Let \( x \) be an observed RGB image, \( y \) be the corresponding ground truth saliency map, and \( z \) be the \( d \)-dimensional vector of latent variables, where \( h \times w \gg d \). In this context, we consider the following generative model to predict a saliency map \( y \) from an image \( x \):

\[
y = T_\theta (x, z) + \epsilon, \quad z \sim p_a (z), \quad \epsilon \sim N (0, \sigma^2 I),
\]

(1)

where \( T_\theta \) is the non-linear mapping process from \( [z, x] \) to \( y \) with parameters \( \theta \), \( p_a (z) \) is the prior distribution with parameters \( \alpha \), and \( \epsilon \sim N (0, \sigma_I^2 I) \) is the observation residual of saliency with \( \sigma_e \) being given. Because of the stochasticity of the latent variables \( z \), the saliency map for a given image \( x \) is also probabilistic. This probabilistic model aligns with the subjective nature of image saliency [57]. To create a generative saliency detection model, one can follow conventional generative models such as GAN [15], VAE [11], and ABP [16] by defining the prior distribution \( p_a (z) \) of the latent variables \( z \) as a simple isotropic Gaussian distribution, resulting in \( p(z) = N (0, I) \), where we drop \( \alpha \) because the prior distribution does not have trainable parameters. The above three generative models differ in the inference process during training. Specifically, VAEs [11] use an additional deep neural network to approximate the posterior distribution of \( z \) for inference, while ABPs [16] directly sample from the true posterior distribution of \( z \) using gradient-based Markov Chain Monte Carlo (MCMC) [19]. In contrast, GANs [15] do not have an explicit inference model but an extra classifier for adversarial learning.

B. Generative Saliency versus Deterministic Saliency

Different from a stochastic model, which produces different outputs given the same input due to random factors, a deterministic model provides a direct mapping between inputs and the outputs. A deterministic saliency detection model directly achieves a mapping function from input image space to the...
output saliency space, given by the following regression model

\[ y = T_θ(x) + ϵ, \quad ϵ \sim \mathcal{N}(0, σ_ϵ^2 I). \]  

(2)

Similarly, \( T_θ \) is the mapping function parameterized by a neural net, and \( ϵ \) represents the observation residual. Compared with the generative saliency formulation in (1), the absence of latent variables \( z \) in the deterministic model shown in (2) does not require dealing with the challenging inference of latent variables, and the estimation of parameters \( θ \) can be easily achieved via supervised learning. The deterministic model corresponds to a point estimation framework, where a fixed prediction is associated with each input image, making it impossible to explore the subjective nature of human saliency. We argue that the inclusion of the extra latent variables \( z \) within the generative saliency detection formulation in (1) leads to a stochastic prediction framework. This makes it natural and effective to represent the distribution of saliency responses over different individuals given an image.

C. Energy-Based Prior versus Gaussian Prior

Although the generative model pipeline is beneficial for uncertainty estimation and representing the subjective nature of saliency, we believe that the effectiveness of the latent space is the core of the generative model. A less informative latent space, such as a simple Gaussian distribution, may fail to accurately reflect the true uncertainty of the saliency, resulting in a less reliable salient object prediction process. Second, an informative prior model, such as an energy-based model, in the latent space can further improve the expressive power of the whole model. Third, the energy function in the EBM prior can be viewed as an objective function or a cost function, which captures regularities and rules in the latent space of the data. Finally, a trainable energy-based prior can be more advantageous than a fixed Gaussian prior as it has the capability to correct the mismatch between the prior and the aggregated posterior resulting from a biased inference process, such as an approximate inference network in VAE or a non-convergent short-run MCMC inference [85] in ABP. In this paper, we depart from the conventional generative approach of assuming a Gaussian distribution for the latent variable and propose the use of an EBM as a prior. The EBM is parameterized with a deep neural network to explore a more informative latent space, resulting in a more robust and reasonable generative saliency prediction framework. Refer to Fig. 2 for a comparison of diagrams between the Gaussian prior and the EBM prior, and Fig. 1 for a qualitative comparison of the prediction results produced by different models.

IV. LEARNING AN ENERGY-BASED PRIOR FOR GENERATIVE SALIENCY DETECTION

A. Model

Given the generative saliency model in (1), following [22], the prior \( p_α(z) \) is not assumed to be a simple isotropic Gaussian distribution as it is GANs [12], VAEs [11, 13] or ABP [16]. Specifically, it is in the form of the energy-based correction or exponential tilting [66] of an isotropic Gaussian reference distribution \( p_0(z) = \mathcal{N}(0, σ^2 I) \), i.e.:

\[ p_α(z) = \frac{1}{Z(α)} \exp \left[ -U_α(z) \right] p_0(z) \]

\[ \propto \exp \left[ -U_α(z) - \frac{1}{2σ^2} ||z||^2 \right], \]

(3)

where \( E_α(z) = U_α(z) + \frac{1}{2σ^2} ||z||^2 \) is the energy function that maps latent variables \( z \) to a scalar, and \( U_α(z) \) is parameterized by a multi-layer perceptron (MLP) with trainable parameters \( α \). The standard deviation \( σ \) is a hyperparameter. \( Z(α) = \int \exp[-U_α(z)]p_0(z)dz \) is the intractable normalizing constant that resolves the requirement for a probability distribution to have a total probability equal to one. \( p_0(z) \) is an informative prior distribution in our model and its parameters \( α \) need to be estimated along with the non-linear mapping function \( T_θ \) from training data.

The mapping function \( T_θ \) encodes the input image \( x \) and then decodes it along with the vector of latent variables \( z \) to the saliency map \( y \), thus, \( p_0(y|x,z) = \mathcal{N}(T_θ(x,z), σ^2 I) \). The resulting generative model is a conditional directed graphical model that combines the EBM prior [22] and the saliency generator \( T_θ \).

B. Maximum Likelihood Learning

The generative saliency prediction model with an energy-based prior, which is presented in (1), can be trained via maximum likelihood estimation. For notation simplicity, let \( β = (θ, α) \). For the training examples \( \{(x_i, y_i), i = 1, \ldots, n\} \), the observed-data log-likelihood function is defined as:

\[ L(β) = \sum_{i=1}^{n} \log p_β(y_i|x_i) = \sum_{i=1}^{n} \log \left[ \int p_β(y_i, z_i|x_i)dz_i \right] \]

\[ = \sum_{i=1}^{n} \log \left[ \int p_α(z_i)p_θ(y_i|x_i, z_i)dz_i \right]. \]

Maximizing \( L(β) \) is equivalent to minimizing the Kullback-Leibler (KL) divergence between the model \( p_β(y|x) \) and the data distribution \( p_{data}(y|x) \). The gradient of \( L(β) \) can be computed.
based on:
\[
\nabla \beta \log p_\beta(y|x) = E_{p_\beta(z|y,x)} [\nabla \beta \log p_\beta(y,z|x)] \\
= E_{p_\beta(z|y,x)} [\nabla \beta (\log p_\alpha(z) + \log p_\theta(y|x,z))],
\]
where the posterior distribution is:
\[
p_\beta(z|y,x) = p_\beta(y,z|x)/p_\beta(y|x) = p_\alpha(z)p_\theta(y|x,z)/p_\beta(y|x).
\]

The learning gradient in (4) can be decomposed into two parts, i.e., the gradient for the energy-based model \(\alpha\):
\[
E_{p_\beta(z|y,x)} [\nabla_\alpha \log p_\alpha(z)] \\
= E_{p_\beta(z|y,x)} [\nabla_\alpha U_\alpha(z)] - E_{p_\beta(z|y,x)} [\nabla_\alpha U_\alpha(z)],
\]
and the gradient for the saliency generator \(\theta\):
\[
E_{p_\beta(z|y,x)} [\nabla_\theta \log p_\theta(y|x,z)] \\
= E_{p_\beta(z|y,x)} \left[ \frac{1}{\sigma^2} (y - T_\theta(x,z)) \nabla_\theta T_\theta(x,z) \right],
\]
(7)
\(\nabla_\theta U_\alpha(z)\) in (6) and \(\nabla_\theta T_\theta(x,z)\) in (7) can be efficiently computed via back-propagation. Both (6) and (7) include intractable expectation terms \(E_{p_\beta(\cdot)}\), which can be approximated by MCMC samples. To be specific, we can use a gradient-based MCMC, e.g., Langevin dynamics, which is initialized with a Gaussian noise distribution \(p_0\), to draw samples from the energy-based prior model \(p_\alpha(z) \propto \exp[-E_\alpha(z)]\) by iterating:
\[
z_{t+1} = z_t - \delta \nabla_\alpha E_\alpha(z_t) + \sqrt{2\delta} \epsilon_t, \quad z_0 \sim p_0(z), \quad \epsilon_t \sim \mathcal{N}(0,I),
\]
and drawing samples from the posterior distribution \(p_\beta(z|y,x)\) by iterating:
\[
z_{t+1} = z_t - \delta \left( \nabla_\alpha U_\alpha(z_t) - \frac{1}{\sigma^2} (y - T_\theta(x,z_t)) \nabla_\theta T_\theta(x,z_t) \right) \\
+ \sqrt{2\delta} \epsilon_t, \quad z_0 \sim p_0(z), \quad \epsilon_t \sim \mathcal{N}(0,I).
\]
(9)
\(\delta\) is the Langevin step size which can be specified independently in (8) and (9). We use \(\{z_{t+}^i\}\) and \(\{z_{t-}^i\}\) to denote, respectively, the samples from the posterior distribution \(p_\beta(z|y,x)\) and the prior distribution \(p_\alpha(z)\). The gradients of \(\alpha\) and \(\theta\) can be computed with \(\{(x_i,y_i)\}\), \(\{z_{i+}\}\) and \(\{z_{i-}\}\) by:
\[
\nabla_\alpha = \frac{1}{n} \sum_{i=1}^n \left[ \nabla_\alpha U_\alpha(z_{i+}^i) - \nabla_\alpha U_\alpha(z_{i-}^i) \right],
\]
\[
\nabla_\theta = \frac{1}{n} \sum_{i=1}^n \left[ \frac{1}{\sigma^2} (y_i - T_\theta(x_i,z_{i+}^i)) \nabla_\theta T_\theta(x_i,z_{i+}^i) \right].
\]
(10)
(11)
We can then update the parameters with \(\nabla_\alpha\) and \(\nabla_\theta\) via the Adam optimizer [86]. We present the full learning and sampling algorithm of our model in Algorithm 1, and the pipeline of saliency detection by a learned model in Algorithm 2.

**C. Analysis**

1) Convergence: Theoretically, when the Adam optimizer of \(\beta = (\theta,\alpha)\) in the learning algorithm converges to a local minimum, it solves the estimating equations \(\nabla_\alpha = 0\) and \(\nabla_\theta = 0\),

**Algorithm 1:** Maximum Likelihood Learning of a Generative Saliency Model With an Energy-Based Prior Model.

**Input:** (1) Training images \(\{x_i\}_n\) with associated saliency maps \(\{y_i\}_n\); (2) Number of learning iterations \(M\); (3) Numbers of Langevin steps for prior and posterior \(\{K^-,K^+\}\); (4) Langevin step sizes for prior and posterior \(\{\delta^-,\delta^+\}\); (5) Learning rates for energy-based prior and saliency generator \(\{\xi_\alpha,\xi_\theta\}\); (6) batch size \(n^*\).

**Output:** Parameters \(\theta\) for the saliency generator and \(\alpha\) for the energy-based prior model
1: Initialize \(\theta\) and \(\alpha\)
2: for \(m \leftarrow 1\) to \(M\)
3: Sample observed image-saliency pairs \(\{(x_i,y_i)\}_n\)
4: For each \((x_i,y_i)\), sample the prior \(z_{t+}^i \sim p_\alpha(z|x_i)\) using \(K^-\) Langevin steps in (8) with a step size \(\delta^-\).
5: For each \((x_i,y_i)\), sample the posterior \(z_{t-}^i \sim p_\beta(z|x_i,y_i)\) using \(K^+\) Langevin steps in (9) with a step size \(\delta^+\).
6: Update energy-based prior using Adam with the gradient \(\nabla_\alpha\) computed in (10) and a learning rate \(\xi_\alpha\).
7: Update the saliency generator using Adam with the gradient \(\nabla_\theta\) computed in (11) and a learning rate \(\xi_\theta\).
8: end for

**Algorithm 2:** Saliency Detection by a Generative Saliency Model With an Energy-Based Prior Model.

**Input:** (1) A testing image \(x^*\); (2) Parameters \(\theta\) for the saliency generator and \(\alpha\) for the energy-based prior model; (3) Number of Langevin steps for the prior model \(K^-\); (4) Langevin step sizes for the prior \(\delta^-\); (5) Number of trials \(J\) for prediction.

**Output:** Saliency prediction \(y^*\) for \(x^*\) and the corresponding uncertainty map \(u^*\)
1: for \(j \leftarrow 1\) to \(J\)
2: Sample latent variables from the prior model \(z_{t+}^j \sim p_\alpha(z)\) using \(K^-\) Langevin steps in (8) with a step size \(\delta^-\).
3: Compute saliency prediction \(y_j = T_\theta(x^*,z_{t+}^j)\).
4: end for
5: Obtain final saliency prediction via \(y^* = \frac{1}{J} \sum_{j=1}^J y_j^*\).
6: Obtain uncertainty as variance of \(J\) trials of predictions:
\[
u^* = \frac{1}{J} \sum_{j=1}^J (y_j - y^*)^2.
\]

which are:
\[
\frac{1}{n} \sum_{i=1}^n E_{p_\beta(z_i|y_i,x_i)} [\nabla_\alpha U_\alpha(z_i)] - E_{p_\alpha(z)} [\nabla_\alpha U_\alpha(z)] = 0,
\]
(12)
\[
\frac{1}{n} \sum_{i=1}^n E_{p_\beta(z_i|y_i,x_i)} \left[ \frac{1}{\sigma^2} (y_i - T_\theta(x_i,z_i)) \nabla_\theta T_\theta(x_i,z_i) \right] = 0.
\]
(13)
The above two equations are the maximum likelihood estimating equations. However, in practice, the Langevin dynamics shown...
in (10) and (11) might not converge to the target distributions due to the use of a small number of Langevin steps (i.e., short-run non-convergent MCMC). As a result, the estimating equations in (13) and (12) will correspond to a perturbation of the MLE estimating equation according to [22], [67], [87]. The learning algorithm can be justified as a Robbins-Monro algorithm, whose convergence is theoretically sound.

2) Accuracy: Compared with the original GAN-based generative framework [15], our model is a likelihood-based generative framework that will not suffer from mode collapse [89]. In comparison with the conventional VAE-based generative framework [11], [13], whose training is also based on likelihood, our MCMC-based maximum likelihood learning algorithm will not encounter the posterior collapse issue that is caused by amortized inference. On the other hand, the variational inference typically relies on an extra inference network for efficient inference of the latent variables given an image and saliency pair, however, the approximate inference might not be able to take the full place of the posterior distribution in practice. To be specific, we use \( q_\theta(z|x, y) \) to denote the tractable approximate inference network with parameters \( \theta \). The variational inference seeks to optimize \( \min_\theta \min_\gamma KL(q_\theta(z|x, y)\|p_\beta(z, y|x)) \), which can be further decomposed into \( \min_\gamma KL(p_{\text{data}}(y|x)\|p_\beta(y|x)) + KL(q_\theta(z|x, y)\|p_\beta(z|x, y)) \). That is, the variational inference maximizes the conditional data likelihood plus a KL-divergence between the approximate inference network and the posterior distribution. Only when the \( KL(q_\theta(z|x, y)\|p_\beta(z|x, y)) \to 0 \), the variational inference will lead to the MLE solution, which is exactly the objective of our model. However, there might exist a gap between them in practice due to the improper design of the inference network. Our learning algorithm is much simpler and more accurate than amortized inference.

3) Computational and Memory Costs Analysis: From the learning perspective, due to the iterative Langevin sampling for the posterior and prior distributions of latent variables, our EBM prior based generative models are more time-consuming for training. For example, the conventional generative models with amortized inference, such as VAE [11], [13], is roughly 1.2 times faster than ours on RGB salient object detection. However, for testing, the extra EBM prior Langevin does not lead to significantly longer inference time, which is comparable with the conventional generative models.

D. Other Variants

We have provided the energy-based prior model in Section IV-A, and introduce the principled maximum likelihood learning strategy in Section IV-B, which can be called EABP, where we infer the latent variables \( z \) by Langevin sampling from its true posterior distribution \( p_\beta(z|y, x) \). In this section, we will utilize adversarial learning and variational inference techniques to develop two variants of our generative saliency model: the EGAN, which is a GAN-based model, and the EVAE, which is an VAE-based model. Both use an energy-based model as a prior distribution.

1) An EBM Prior for GAN-Based Generative Saliency: We can train the generative saliency prediction model with an energy-based prior through adversarial learning. This can be achieved by introducing an extra fully convolutional conditional discriminator [61], denoted as \( D_\gamma \), where \( \gamma \) refers to the parameters of the discriminator. Following the conventional conditional adversarial learning pipeline, the objective of our framework can be expressed as:

\[
L_{\text{adv}}(\gamma, \theta) = \mathbb{E}_{(x, y) \sim p_{\text{data}}(x, y)} \left[ \log D_\gamma(y, x) \right] + \mathbb{E}_{x \sim p_{\text{data}}(x), z \sim p_\alpha(z)} \left[ \log (1 - D_\gamma(T_\theta(x, z), x)) \right],
\]  

(14)

where \( p_{\text{data}}(x, y) \) is the joint data distribution, \( p_{\text{data}}(x) \) is the marginal data distribution, and \( p_\alpha(z) \) is the prior distribution of the latent variables \( z \), which is usually defined as a known Gaussian distribution in the conventional GAN framework but now is a trainable energy-based distribution in our model.

Previous approaches have found it beneficial to combine the GAN objective with a more traditional reconstruction loss, such as L2 [90] or L1 [91] distance. While the discriminator’s role remains unchanged, the generator aims to not only fool the discriminator but also produce outputs that are closer to the ground truth. In our case, where we are predicting saliency, the reconstruction loss can be calculated using a binary cross-entropy loss or a structure-aware loss [2] between the generated outputs and the ground truth outputs, that is:

\[
L_{\text{rec}}(\theta) = \mathbb{E}_{(x, y) \sim p_{\text{data}}(x, y), z \sim p_\alpha(z)} \left[ \ell(T_\theta(x, z), y) \right],
\]

(15)

where \( \ell(y', y) \) calculates the cross-entropy between \( y' \) and \( y \). Our final objective to learn the saliency generator \( \theta \) and the discriminator \( \gamma \) is given by:

\[
\min_\theta \max_\gamma L_{\text{adv}}(\gamma, \theta) + \lambda L_{\text{rec}}(\theta),
\]

(16)

where the hyper-parameter \( \lambda \) is empirically set to be 0.1 for stable training. Both (14) and (15) require sampling from the prior distribution to generate saliency outputs. We follow Langevin dynamics in (8) to draw samples from \( p_\alpha(z) \). The update of the parameters \( \alpha \) of the energy-based prior model follows (10), in which we need to sample from the EBM prior and the posterior distribution \( p_\beta(z|y, x) \).

We highlight a significant issue with conventional conditional GANs, namely that the distribution of the latent space is independent of training data. This makes it less effective in representing the latent space of the data distribution. Instead of assuming the prior to be an isotropic Gaussian distribution, we adopt an EBM prior model in EGAN and train it through empirical Bayes. Additionally, in contrast to the EABP, the EGAN recruits an additional assisting network \( D \) and trains the parameters of the saliency generator \( \gamma \) by an adversarial loss without relying on inference. However, the inference step is still a necessary part of training the EBM prior.

At a testing stage, we can follow the same testing pipeline as in Algorithm 2, where we first sample latent variables from the EBM prior through Langevin dynamics and then transform the latent variables and the input image testing by the saliency generator to produce a saliency prediction.

2) An EBM Prior for VAE-Based Generative Saliency: We can also train the generative saliency prediction model with
an energy-based prior through variational inference [11], [13], [23]. This can be achieved by introducing an extra tractable inference model \( q_\phi(z | y, x) \), with parameters \( \phi \), to approximate the intractable posterior distribution \( p_\beta(z | y, x) \). Following the reparameterization trick [11], we design the inference model 
\[
q_\phi(z | y, x) = \mathcal{N}(\mu_\phi(y, x), \text{diag}(\sigma_\phi(y, x))),
\]
where \( \mu_\phi(y, x) \) and \( \sigma_\phi(y, x) \) are \( d \)-dimensional outputs of encoding bottom-up networks that take \( x \) and \( y \) as input.

Given an image \( x \) with its saliency map \( y \), the log-likelihood \( \log p_\beta(y | x) \) is lower bounded by the evidence lower bound (ELBO), which is:

\[
\text{ELBO}(\beta, \phi, x, y) = \log p_\beta(y | x) - \mathbb{D}_{KL}(q_\phi(z | y, x) \| p_\beta(z | y, x)) = \mathbb{E}_{q_\phi(z | y, x)}[\log p_\beta(y | z, x)] - \mathbb{D}_{KL}(q_\phi(z | y, x) \| p_\alpha(z | x)),
\]

where we assume \( z \) and \( x \) are independent, so that we can simply set the prior model \( p_\alpha(z | x) = p_\alpha(z) \propto \exp(-U_\alpha(z))p_0(z) \).

As to (17), the first term is the reconstruction loss and the second one is the Kullback-Leibler divergence between the inference network and the EBM prior distribution.

For the prior model \( \alpha \), the learning gradient is:

\[
\nabla_\alpha \text{ELBO}(\beta, \phi, x, y) = \mathbb{E}_{p_\alpha(z)}[\nabla_\alpha U(z)] - \mathbb{E}_{q_\phi(z | y, x)}[\nabla_\alpha U(z)],
\]

where \( \mathbb{E}_{p_\alpha(z)} \) is approximated by MCMC samples from the EBM prior \( p_\alpha(z) \), and \( \mathbb{E}_{q_\phi(z | y, x)} \) is approximated by samples from the inference network.

For the bottom-up inference model \( \phi \) and the top-down saliency generator \( \psi \), let \( \psi = \{\theta, \phi\} \) and the learning gradients of \( \psi \) are given by:

\[
\nabla_\psi \text{ELBO}(\beta, \phi, x, y) = \nabla_\psi \mathbb{E}_{q_\phi(z | y, x)}[\log p_\beta(y | z, x)]
- \nabla_\psi \mathbb{D}_{KL}(q_\phi(z | y, x) \| p_\beta(z))
- \nabla_\psi \mathbb{E}_{q_\phi(z | y, x)}[U_\alpha(z)],
\]

where the first term is the derivative of the reconstruction loss and the second term is the derivative of the Kullback-Leibler divergence between the inference distribution and the Gaussian reference distribution \( p_0(z) \) in the EBM prior model.

\[
\mathbb{D}_{KL}(q_\phi(z | y, x) \| p_\beta(z))
\]

is tractable. The third term is approximated by the samples from the inference network. In contrast to the conventional VAE loss [11] using a Gaussian prior, (19) has an extra term, which is the third term involving the energy function \( U_\alpha(z) \).

E. Network Architectures

1) Saliency Generator: We utilize the Swin transformer backbone [44] to design our saliency generator and perform ablation studies with CNN backbones in Section V-C. As depicted in Fig. 3, our saliency generator takes a three-channel image \( x \) and the latent variables \( z \) as inputs and generates a one-channel saliency map \( T_\theta(x, z) \).

![Fig. 3. Architecture of the saliency generator for saliency generation.](image)

Our saliency generator includes two main modules: the “Transformer Encoder” module and the “Feature Aggregation” module. The former module takes \( x \) as input and produces a set of feature maps \( \{f_l\}_{l=1}^5 \) of channel sizes 128, 256, 512, 1024, and 1024, respectively. The latter module takes \( \{f_l\}_{l=1}^5 \) and the latent variables \( z \) as inputs to generate the saliency prediction \( y \). To achieve this, we first apply a multi-scale dilated convolutional layer [92] to each \( f_l \) to reduce the channel dimension of each level feature \( f_l \) to a channel size of \( C = 32 \). This results in a new set of feature maps \( \{f_l\}_{l=1}^5 \).

Next, we spatially replicate the vector \( z \) and concatenate it channel-wise with \( f_5 \), followed by a \( 3 \times 3 \) convolutional layer that produces a feature map \( F_5 \) with the same number of channels as \( f_5 \). Finally, we sequentially concatenate feature maps from high to low levels via feature aggregation, i.e., from \( l = 4 \) to 1, we compute \( F_l = \text{Conv}_{3x3}(\text{M}(\text{Concat}(f_l, F_{l+1}, \ldots, F_5))) \), where \( \text{Conv}_{3x3}(-) \) is a \( 3 \times 3 \) convolutional layer that reduces the channel dimension to 32, \( \text{M}(-) \) is the channel attention module [93], and \( \text{Concat}(-) \) is the channel-wise concatenation operation.

It is important to note that we upsample the higher-level feature map to the same spatial size as that of the lower-level feature map before the concatenation operation. We then feed \( F_1 \) to a \( 3 \times 3 \) convolutional layer to obtain the one-channel saliency map \( T_\theta(x, z) \).

2) Energy-Based Prior Model: We develop an energy-based model to represent the prior distribution of the latent variables \( z \), utilizing a multilayer perceptron (MLP) to parameterize the energy function \( U_\alpha(z) \). Our MLP comprises three fully connected layers that map the latent variables \( z \) to a scalar. The feature map sizes for the MLP’s different layers are \( C_1, C_2, \) and \( 1 \), respectively. For the size of the EBM prior, we will denote it as \( C_5 \) and set it to 60 for our experiment. We use GELU [94] activation after each layer except the last one.

3) Discriminator in the EGAN: The discriminator \( D_\gamma \) in EGAN consists of five \( 3 \times 3 \) convolutional layers that take as input the concatenation of the image \( x \) and the prediction \( T_\theta(x, z) \) (or the ground truth map \( y \)) and output an all-zero feature map \( 0 \) (or an all-one feature map \( 1 \)). The first four convolutional layers have a channel size of \( C_d = 64 \), and the last convolutional layer has a channel size of \( 1 \). To increase the receptive field, we apply a stride of 2 during the convolution operation in the 1st, 3rd, and 5th layers. We also apply batch normalization and Leaky RELU after the first four convolutional layers.

4) Inference Model in the EVAE: The inference model for approximate posterior distribution \( q_\phi(z | x, y) \) is composed of five convolutional layers and two fully connected layers. The network maps the concatenation of the RGB image \( x \) and its saliency map \( y \) to feature maps with channel sizes of \( C_1, 2 \times C_1, 4 \times C_1, 8 \times C_1, \) and \( 8 \times C_1 \) through five convolutional layers using a kernel size of \( 4 \times 4 \) and a stride of 2. We set \( C_1 = 32 \). The subsequent two fully connected layers map the \( 8 \times C_1 \)-channel feature map to the mean \( \mu \) of the latent variables \( z \) and
its standard deviation \( \sigma \). We perform batch normalization and Leaky RELU after each convolution operation. The inference distribution of the latent variables \( z \) is then obtained via the reparameterization trick: \( z = \mu + \epsilon \times \sigma \), where \( \epsilon \sim \mathcal{N}(0, 1) \).

V. EXPERIMENTAL RESULTS

A. Setup

1) Tasks, Datasets, and Evaluation Metrics: In this paper, we address two common static image-based saliency detection tasks: (i) RGB saliency detection and (ii) RGB-D saliency detection. For RGB saliency detection, we train models on DUTS training set [95], which consists of 10,553 images with the corresponding pixel-wise annotations. We then test model performance on six benchmark RGB saliency detection testing datasets, including the DUTS testing set, ECSSD [96], DUT [97], HKU-IS [98], PASCAL-S [99] and the SOD testing dataset [100]. For RGB-D saliency detection, our aim is to identify the most visually distinctive objects in a scene from the provided RGB and depth data. We follow the conventional training setting, in which the training set is a combination of 1,485 images from NJU2K dataset [106] and 700 images from NLP dataset [107]. We test the trained models on NJU2K testing set, NLP testing set, LFSKD [108], DES [109], SSB [110] and SIF [111] testing set. We employ four evaluation metrics, including Mean Absolute Error \( M \), Mean F-measure (\( F_\beta \)), Mean E-measure (\( E_\beta \)) [112] and S-measure (\( S_\alpha \)) [113], to measure the performance of both RGB saliency detection models and RGB-D saliency detection models.

2) Implementation Details: Our generative model is built upon the Swin transformer [44] and we use it as the backbone of the encoder in our framework. The Swin backbone is initialized with the parameters pretrained on the ImageNet-1K [114] dataset for image classification. The other parameters of the newly added components, including the decoder part, the MLP of the energy based prior model, the fully convolutional discriminator within EGAN and the prior along with the inference net within EVAE will be randomly initialized from the Gaussian distribution \( \mathcal{N}(0, 0.01) \). Empirically we set the number of dimensions \( d \) of the latent variables \( z \) as \( d = 32 \). We set \( \sigma_r = 1 \) in (1) and \( \sigma_z = 1 \) in (3). We resize all the images and the saliency maps to the resolution of \( 384 \times 384 \) pixels to fit the Swin transformer. The maximum epoch is 50 for RGB saliency detection and 100 for RGB-D saliency detection. The initial learning rates of the saliency generator, the discriminator and the EBM are \( 2.5 \times 10^{-5}, 1 \times 10^{-5} \) and \( 1 \times 10^{-4} \) respectively. For both the prior and posterior Langevin processes, we set the numbers of Langevin steps \( K^- = K^+ = 6 \). The Langevin step sizes of the prior and posterior are \( \delta^- = 0.4 \) and \( \delta^+ = 0.1 \). For RGB-D saliency detection, we perform early fusion by concatenating the RGB image and the depth data at the input layer. We then feed this concatenated data to a 3 \( \times \) 3 convolutional layer before passing it to the Swin encoder [44]. As to the computational cost of EBM prior-based models, the complete training process requires roughly 17 hours, using a batch size of \( n' = 10 \), on a single NVIDIA GTX 2080Ti GPU for an RGB saliency prediction model. For a model related to RGB-D saliency, the training time is about 12 hours. During testing, our model is capable of processing 15 images per second with one time of sampling from the EBM prior distribution.

B. Performance Comparison

1) RGB SOD Performance Comparison: We compare the proposed EBM prior-based generative saliency prediction models, including EABP, EVAE, and EGAN, with the state-of-the-art (SOTA) RGB SOD models, and show performance comparison in Table I. All of these models utilize the Swin transformer backbone [44]. Notably, our generative models’ performance is reported by averaging the results of ten prediction generation trials. As Table I illustrates, our proposed frameworks outperform the state-of-the-art models, exhibiting the superiority of generative saliency prediction models.

Furthermore, we provide visualizations of the predictions generated by our models and two state-of-the-art (SOTA) models in Fig. 4 to analyze the effectiveness of our proposed framework. Specifically, for the input RGB image shown in the first row, the salient foreground is dark with poor illumination. The SOTA models produce numerous false positives, while our models accurately localize the entire salient foreground. The salient foreground of the second RGB image in the second row is in a cluttered background, which is commonly considered a complex case for saliency detection. Existing techniques struggle to distinguish the salient region from the complex background. However, the precise predictions of all three of our generative models provide further evidence of the superiority of our generative frameworks.

2) RGB-D SOD Performance Comparison: We compare the performance of our RGB-D saliency detection models, including EABP, EVAE, and EGAN, with the state-of-the-art (SOTA) RGB-D SOD models in Table II. Our “early-fusion network architecture” show a clear advantage compared to existing techniques on five out of six benchmark testing datasets. However, we observe that our generative models perform worse on the LFSKD dataset [108] than the state-of-the-art RGB-D saliency prediction models that use a cross-level-fusion network architecture, namely DSA2F [40] and VST [24]. LFSKD [108] is a small RGB-D saliency detection dataset consisting of 100 RGB-D images. To better understand the limitations of our models on this dataset, we visualize some examples where the VST [24] outperforms our model (EGAN in this case) in Fig. 5.

In our analysis, we observe that the appearance information from the RGB image in the first row contributes more to saliency detection, whereas equal contribution of appearance and geometric information are observed for the second example. However, the geometric information dominates RGB-D saliency for the third example. Our failure cases in Fig. 5 further verify the necessity of effective cross-level feature fusion. Although our model performs early-fusion, it may work inferiorly compared to state-of-the-art cross-level fusion models [24], [40]. To address this issue, we plan to introduce an effective cross-level fusion strategy within our generative saliency detection framework.
3) Uncertainty Comparison: As a generative learning framework, we further compare with existing generative saliency detection models [9] with respect to the quality of uncertainty maps in Fig. 6. UCNet [9] is a conditional variational auto-encoder (CVAE) [13] for RGB-D saliency detection. Different from UCNet [9] which uses a Gaussian latent variable assumption, our three new generative models (i.e., EGAN, EVAE and EABP) use an EBM prior for the latent space. The inference process of the latent variables is performed using gradient-based MCMC [19] in EABP, which helps to avoid the posterior collapse issue [14] within the conventional CVAE-based framework. Fig. 6 shows that our three models can obtain more accurate saliency prediction performance and more reliable uncertainty maps than the baseline UCNet.

C. Backbone Analysis

As most saliency detection models use CNN backbones, we test our models using two widely used CNN backbones (ResNet50 and VGG16), and show performance for both RGB and RGB-D saliency detection in Tables III and IV, respectively.
TABLE III

| Backbone | Method | DUTS [95] | ECSSD [96] | DUT [97] | HKU-IS [98] | PASCAL-S [99] | SOD [100] |
|----------|--------|-----------|------------|----------|-------------|---------------|---------|
| ResNet50 | EGAN   | 881.81     | 838.13     | 915.91   | 942.03      | 829.72        | 862.68  |
|          | EVAE   | 879.83     | 833.91     | 915.90   | 942.03      | 820.72        | 848.66  |
|          | EABP   | 877.83     | 831.91     | 916.91   | 941.03      | 825.74        | 855.66  |
| EABP_A   |        | 896.89     | 859.93     | 925.92   | 947.03      | 840.78        | 871.05  |

TABLE IV

| Backbone | Method | NU2K [106] | SSB [110] | DES [109] | NLP17 [107] | LFPD [108] | SIP11 [109] |
|----------|--------|------------|-----------|-----------|-------------|-------------|-------------|
| ResNet50 | EGAN   | 910.89     | 939.83    | 905.89    | 936.83      | 906.89       | 998.83     |
|          | EVAE   | 909.89     | 938.83    | 905.89    | 936.83      | 905.89       | 997.83     |
|          | EABP   | 907.89     | 937.83    | 905.89    | 936.83      | 904.89       | 996.83     |
| EABP_A   |        | 911.89     | 939.83    | 905.89    | 936.83      | 906.89       | 998.83     |

Fig. 5. An illustration of examples from LFSD dataset [108], in which the state-of-the-art model, VST [24], that uses a cross-level-fusion network architecture, outperforms our EGAN framework that adopts an early-fusion network architecture.

We observe that our generative models using the Swin transformer backbone [44] perform better than those using conventional CNN backbones for both RGB and RGB-D saliency detection tasks, which indicates the superiority of the Swin transformer backbone for these tasks. Additionally, we find that the ResNet50 backbone perform slightly better than the VGG16 backbone. During the training of the two CNN backbone-based models, we use the same initial learning rates and maximum number of learning epochs. However, as the ResNet50 backbone converges faster than the VGG16 backbone, the relatively worse performance of the VGG16 backbone-based generative models may be due to less thorough training. We plan to investigate this in future research.
In our generative models, we do not use widely adopted data augmentation techniques, which are common in existing literature. For a fair comparison, we also introduce multi-scale training to the proposed “EABP” models to achieve data augmentation, leading to “EABP_A” models shown in both Tables III and IV. This involves performing training on three scales (i.e., 0.75, 1, 1.25) of the original training images. However, as the Swin transformer backbone [44] can only take a fixed input size, we can not perform multi-scale training for the Swin-related generative models. The improved performance of our models with the data augmentation technique highlights the effectiveness of data augmentation for saliency detection models.

Tables III and IV demonstrate the advantages of using the Swin transformer as a backbone for saliency detection. To further analyze the difference in uncertainty modeling abilities between transformer-based and CNN backbone-based models, we visualize the uncertainty maps and prediction gaps of Swin transformer-based model and ResNet50 backbone-based model in Fig. 7. We only show the “EABP” model of each backbone for simplicity. The qualitative prediction results shown in Fig. 7 demonstrate the superior performance of the transformer backbone-based generative model, which can be attributed to the long-range dependency modeling ability of the vision transformer structure. Moreover, The difference map between the saliency prediction shown in the fifth column reveal the performance gap between the transformer and the CNN backbones, and the uncertainty maps of models based on both backbones are generally reasonable. Thus, we can conclude that although the performance of transformer and CNN backbones differ, their uncertainty modeling abilities within our EBM prior-based generative models are useful and reasonable.

D. Ablation Study

We’ve conducted additional experiments. Unless otherwise specified, the experiments pertain to RGB saliency detection.

1) Analyzing Deterministic Component: We first investigate the deterministic component of our framework. We first design a deterministic base model for saliency detection by feeding the backbone features directly to the decoder to generate a one-channel saliency map. We design three base models with ResNet50, VGG16 and Swin transformer backbones respectively. The performance of these models is presented as “Base” in each “Backbone” category of Table V, which clearly demonstrates the advantage of using transformer backbones for saliency detection. The superior performance of the Swin transformer backbone [44] compared to ResNet50 and VGG16 can be attributed to its massive number of parameters. Specifically, the Swin transformer contains 90 M parameters for the “Base” model, which is almost double the number of parameters of ResNet50 (52 M) and triple the number of parameters of VGG16 (29 M). The increased number of parameters allows the Swin transformer to capture more complex and high-level features, resulting in improved performance.

2) Analyzing Stochastic Component: Building upon the deterministic models, we introduce Gaussian latent variables and design the corresponding stochastic models, namely GAN [15], VAE [11], and ABP [16]. Their performance is indicated as “GAN”, “VAE” and “ABP”, respectively, in Table V. For stochastic models using VGG16 and ResNet50 backbones, we observe similar or inferior performance compared to the base models. However, for the Swin transformer-based models, we find comparable or improved performance of the stochastic models (especially “GAN” and “VAE”) compared to the base model. We argue that the reason for this may be the different capacity of each backbone model. The latent variables in the larger network can be more expressive to capture the underlying randomness of the saliency maps. In contrast, for a smaller network, the latent variables are more sensitive to the output, and the model requires more careful tuning to achieve a good performance. In this study, we use the same hyper-parameters for the three generative models with different backbones, including the number of dimensions of the latent space and the
TABLE V
AN ABLATION STUDY THAT ANALYZES THE DETERMINISTIC AND STOCHASTIC COMPONENTS OF OUR PROPOSED FRAMEWORK FOR RGB SALIENT OBJECT DETECTION

| Backbones | Method | ECSSD [96] | DUTS [95] | HKU-IS [98] | PASCAL-S [99] | SOD [100] |
|-----------|--------|------------|------------|-------------|----------------|-------------|
| ResNet50  |        |            |            |             |                |             |
| Base      |        |            |            |             |                |             |
| GAN       | 0.83   | 0.83       | 0.84       | 0.84        | 0.83           | 0.85        |
| VAE       | 0.85   | 0.85       | 0.85       | 0.85        | 0.85           | 0.85        |
| ABP       | 0.87   | 0.87       | 0.87       | 0.87        | 0.87           | 0.87        |
| VGG16     |        |            |            |             |                |             |
| Base      |        |            |            |             |                |             |
| GAN       | 0.87   | 0.87       | 0.87       | 0.87        | 0.87           | 0.87        |
| VAE       | 0.88   | 0.88       | 0.88       | 0.88        | 0.88           | 0.88        |
| ABP       | 0.89   | 0.89       | 0.89       | 0.89        | 0.89           | 0.89        |
| Swin      |        |            |            |             |                |             |
| Base      |        |            |            |             |                |             |
| GAN       | 0.91   | 0.91       | 0.91       | 0.91        | 0.91           | 0.91        |
| VAE       | 0.90   | 0.90       | 0.90       | 0.90        | 0.90           | 0.90        |
| ABP       | 0.89   | 0.89       | 0.89       | 0.89        | 0.89           | 0.89        |

Fig. 8. Comparison of RGB saliency detection performance (MAE) among the base models (“Base”), the conventional generative models (“GAN”, “VAE” and “ABP”) and the proposed EBM prior-based generative models (“EGAN”, “EVAE” and “EABP”).

3) Deterministic Vs. Stochastic Using Gaussian Prior Vs. Stochastic Using EBM Prior: We compare deterministic models (“Base”), stochastic models using a Gaussian prior (“GAN”, “VAE”, and “ABP”), and stochastic models using an EBM prior (“EGAN”, “EVAE”, and “EABP”) in terms of MAE, and show the results in Fig. 8. We observe similar performance of the models on easier testing datasets, such as ECSSD [96] and HKU-IS [98]. However, for harder datasets, such as DUTS [95] and SOD [100], we observe obvious differences in terms of MAE among the models. This is consistent with the fact that observers tend to generate relatively consistent salient objects for easier samples, resulting in lower prediction variance. However, complex images have a higher degree of inherent uncertainty, thus leading to various uncertainty results generated by different generative models. To qualitatively demonstrate the uncertainty modeling abilities of the those generative models, we generate uncertainty maps of both conventional generative models and our EBM prior-based generative models with the Swin transformer backbone, which are shown in Fig. 1. The uncertainty maps are obtained by computing the variance of multiple stochastic prediction outputs, which visualize the pixel-wise confidence of the model for saliency prediction.

4) Applying the Proposed Stochastic Framework to SOTA Saliency Detection Model: As general generative models, we apply our proposed EBM prior-based generative frameworks to the existing SOTA RGB and RGB-D saliency detection models. Especially, we choose VST [24] as the base model as it provides models for both RGB and RGB-D saliency detection tasks. The performance is shown in Table VI, where we also include performance of the original deterministic VST model [24] for reference. Specifically, the generative VST models are achieved by following the same manner as our generative models presented in Tables III and IV. Table VI shows similar deterministic performance of the EBM prior-based generative VST [24] compared with the deterministic VST [24], where the deterministic performance is achieved by sampling ten iterations from the EBM prior model, and performance of the mean predictions are reported. Table VI indicates that the EBM prior-based generative VST has comparable deterministic performance with the deterministic VST [24]. The deterministic performance is obtained by sampling ten latent vectors from the EBM prior model, and then computing the average performance over the ten trials of predictions. We further visualize the generated uncertainty maps of the stochastic RGB saliency detection models in Fig. 9. The visualizations demonstrate that while our generative VST saliency models may not significantly improve the VST performance, they are able to generate uncertainty maps that help explain the prediction behavior of the models and reveal the stochastic nature of salient object prediction in humans.

5) Hyperparameter Analysis: The main hyperparameters in our framework include the number of Langevin steps $K$, the Langevin step size $\delta$, the number of dimensions of the latent space $d$. We have two sets of hyperparameters $\{\delta^-, K^-\}$ and $\{\delta^+, K^+\}$ of the Langevin dynamics for sampling from the
TABLE VI
APPLYING THE PROPOSED GENERATIVE FRAMEWORKS (INCLUDING EGAN, EVAE, AND EABP) TO STATE-OF-THE-ART DETERMINISTIC VST MODELS FOR RGB AND RGB-D SALIENT OBJECT DETECTION

| Task | Method | DUTS [95] | ECSSD [96] | DUT [97] | HKU-IS [98] | PASCAL-S [99] | SOD [100] |
|------|--------|-----------|------------|----------|-------------|--------------|------------|
| RGB  | VST [24] | 0.894 ± 0.028 | 0.918 ± 0.037 | 0.924 ± 0.047 | 0.928 ± 0.043 | 0.986 ± 0.058 | 0.963 ± 0.069 |
|      | EGAN   | 0.895 ± 0.049 | 0.916 ± 0.038 | 0.950 ± 0.047 | 0.952 ± 0.053 | 0.986 ± 0.069 | 0.963 ± 0.079 |
|      | EVAE   | 0.893 ± 0.041 | 0.914 ± 0.039 | 0.931 ± 0.047 | 0.931 ± 0.047 | 0.973 ± 0.059 | 0.952 ± 0.070 |
|      | EABP   | 0.895 ± 0.038 | 0.916 ± 0.036 | 0.935 ± 0.044 | 0.947 ± 0.059 | 0.973 ± 0.059 | 0.952 ± 0.070 |

For the number of Langevin steps, we empirically set $K = 5$ to achieve a trade-off between the training efficiency and the model performance, as more Langevin steps will lead to longer training time but better convergence. Additionally, we investigate the influence of the number of latent dimensions by varying $d = \{8, 16, 32, 64\}$, and observe comparable performance among different choices of $d$. We set $d = 32$ in our paper.

6) Analyzing Model Performance Over Percentage of Salient Foreground: We further analyze model performance with respect to the percentage of the salient foreground of the whole image, which can be one of the main factors that lead to various performance of different saliency detection models. We consider ReNet50, VGG16 and Swin backbones. We first select one backbone as a baseline method, and then compute the S-measure performance and the salient foreground percentage for each image on the DUT testing dataset [97]. Then we keep those image examples with S-measure scores less than 0.5, indicating failure cases. We test other two backbones on these failure cases and plot curves of the S-measure performance over percentage of salient foreground for all three backbones.

![Visual comparison of VST model [24] and its generative variants, including EGAN, EVAE, and EABP.](image)

Fig. 9. Visual comparison of VST model [24] and its generative variants, including EGAN, EVAE, and EABP. Each row presents one example of salient object detection. For each example, the first column shows a testing image and the second column shows the ground truth saliency map. The 3rd, 4th, 6th, and 8th columns, respectively, display the prediction outputs of VST, EGAN, EVAE and EABP. For each prediction output of the generative models, we also display the corresponding uncertainty map.

![Model performance w.r.t. percentage of the salient foreground.](image)

Fig. 10. Model performance w.r.t. percentage of the salient foreground.

In Fig. 10, each sub-figure corresponds to one base backbone we select and displays S-measure curves over the percentage of salient foreground.
salient foreground for all the three backbones. For example, in the “ResNet50” sub-figure, we use testing examples where the S-measure performance scores of the “ResNet50” base model is less than 0.5, and plot the S-measure curves for all backbone candidates by using those selected samples. The “ResNet50” sub-figure shows how other backbones (i.e., “VGG16” and “Swin transformer”) perform on the failure examples of the “ResNet50”.

Fig. 10 shows that when the performance of the two CNN backbone-based models are unsatisfactory, the Swin transformer backbone-based model can still perform consistently well across the different foreground scales, which can be observed in the top and the middle sub-figures of Fig. 10. We can also observe CNN backbone-based models perform well on examples on which the Swin transformer backbone performs poorly, thus motivating us to combine both CNN backbone and Swin transformer backbone in the future to effectively explore their individual advantages.

It is worth noting that the range of foreground scales for the failure cases of the two CNN backbone-based models is larger than that of the Swin transformer backbone-based model. Specifically, the foreground scale ranges are [0, 0.7] for ResNet50, [0, 0.9] for VGG16, and [0, 0.33] for Swin transformer. This suggests that the Swin transformer-based model is more capable of handling smaller salient objects than the other two CNN backbone-based models, which is consistent with the Swin transformer’s powerful long-range dependency modeling ability. However, we also observe that all three backbones still struggle with images containing tiny salient objects, as their performance in the range of [0, 0.1] is highly unstable. Our future work will be focused on addressing the challenging task of detecting tiny salient objects.

VI. CONCLUSION AND DISCUSSION

In this paper, we study generative modeling and learning of RGB and RGB-D salient object detection. We start by defining a conditional probability distribution of saliency maps given an input image by a top-down latent variable generative framework, in which the non-linear mapping from image domain to saliency domain is parameterized by a vision transformer network and the prior distribution of the low-dimensional latent space is represented by a trainable energy-based model. Instead of using amortized inference and sampling strategies, we learn the model by the MCMC-based maximum likelihood, where the Langevin sampling is used to evaluate the intractable posterior and prior distributions of the latent variables for calculating the learning gradients of the model parameters. With the informative energy-based prior and the expressive top-down vision transformer network, our model can achieve both accurate predictions and meaningful uncertainty maps that are consistent with the human perception. We further propose two variants of the EBM prior-based generative framework: VAE-based and GAN-based generative saliency models using EBM as a prior. To extensively analyze the superiority of the proposed generative framework, we conduct a series of experiments for an ablation study. We also apply our frameworks to existing SOTA deterministic saliency detection models to further verify the generalization ability of our frameworks.

From the machine learning perspective, our model is a top-down deep conditional generative model using EBM as a prior. The MLE learning algorithm derived from the proposed model is based on MCMC inference for the posterior and MCMC sampling for the prior, which makes our framework more natural, principled, and statistically rigorous than others. On the other hand, the traditional adversarial learning and variational learning techniques can be easily adapted to the proposed EBM prior-based generative model, which can bring in benefits of good performance and fast inference respectively. The proposed frameworks are not only beneficial for saliency prediction but can also be applied to other conditional learning scenarios such as semantic segmentation, image-to-image translation, and more. For instance, we can utilize the same modeling approach to construct a conditional distribution of the target image given the source image for stochastic image-to-image translation or a conditional distribution of semantic segmentation map given the RGB image for generative semantic segmentation. The only variation in these generalizations lies in the network architecture design to represent various modalities. Therefore, the proposed generative model and learning algorithm are generic and versatile.

From the computer vision perspective, our model with a special network design to handle saliency prediction is a new member of the family of saliency prediction methods. In comparison with the traditional discriminative saliency prediction methods, our generative method is natural because it models the saliency prediction as a conditional probability distribution. Further, the EBM prior leads to more reliable uncertainty modeling compared with the conventional generative models.

The use of the latent space energy-based prior model poses a challenging problem for MCMC sampling from an unnormalized probability distribution. Our model requires an iterative MCMC to evaluate the EBM, but MCMC via Langevin dynamics is time-consuming. To strike a balance between accurate evaluation and efficient sampling, we may use short-run MCMC chains in practice, which may lead to non-convergent MCMC sampling issues. Therefore, further investigation is expected to understand the learning behaviors of our proposed algorithms in the context of non-convergent short-run MCMC sampling.

REFERENCES

[1] J. Zhang, N. B. Jianwen Xie, and P. Li, “Learning generative vision transformer with energy-based latent space for saliency prediction,” in Proc. Conf. Neural Inf. Process. Syst., 2021.
[2] J. Wei, S. Wang, and Q. Huang, “F3Net: Fusion, feedback and focus for salient object detection,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 12321–12328.
[3] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, “Label decoupling framework for salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 13025–13034.
[4] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, “BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network,” in Eur. Conf. Comput. Vis., 2020, pp. 275–292.
[5] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, “JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3052–3062.
[6] H. Chen and Y. Li, “Progressively complementarity-aware fusion network for RGB-D salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3051–3060.

[7] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, “Weakly-supervised salient object detection via scribble annotations,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 12543–12552.

[8] Z. Wu, L. Su, and Q. Huang, “Stacked cross refinement network for edge-aware salient object detection,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 7264–7273.

[9] J. Zhang et al., “UC-net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 8582–8591.

[10] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?,” 2017, arXiv:1703.04977.

[11] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” in Proc. Int. Conf. Mach. Learn., 2014.

[12] J. Goodfellow et al., “Generative adversarial nets,” in Proc. Conf. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[13] K. Sohn, H. Lee, and X. Yan, “Learning structured output representation using deep conditional generative models,” in Proc. Conf. Neural Inf. Process. Syst., 2015, pp. 3483–3491.

[14] J. He, D. Spokony, G. Neubig, and T. Berg-Kirkpatrick, “Lagging inference networks and posterior collapse in variational autoencoders,” in Proc. Int. Conf. Mach. Learn., 2019.

[15] I. Goodfellow et al., “Generative adversarial nets,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2672–2680.

[16] T. Han, Y. Lu, S.-C. Zhu, and Y. N. Wu, “Alternating back-propagation for generator network,” in Proc. AAAI Conf. Artif. Intell., 2017, pp. 1976–1982.

[17] J. Xie, R. Gao, Z. Zheng, S.-C. Zhu, and Y. N. Wu, “Learning dynamic generator model by alternating back-propagation through time,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 5498–5507.

[18] J. S. Liu, Monte Carlo Strategies in Scientific Computing, Berlin, Germany: Springer, 2008.

[19] R. M. Neal, “MCMC using Hamiltonian dynamics,” in Handbook of Markov Chain Monte Carlo, Boca Raton, FL, USA: Chapman & Hall/CRC Press.

[20] M. Wellings and Y. W. Teh, “Bayesian learning via stochastic gradient Langevin dynamics,” in Proc. Int. Conf. Mach. Learn., 2011, pp. 681–688.

[21] K. A. Dubey, S. J. Reddi, S. A. Williamson, B. Poczos, A. J. Smola, and E. P. Xing, “Variance reduction in stochastic gradient Langevin dynamics,” in Proc. Conf. Neural Inf. Process. Syst., 2016, pp. 1154–1162.

[22] B. Pang, T. Han, E. Nijkamp, S. Zhu, and Y. N. Wu, “Learning latent space energy-based prior model,” in Proc. Conf. Neural Inf. Process. Syst., 2020.

[23] B. Pang and Y. N. Wu, “Latent space energy-based model of symbol-vector coupling in next generation classification,” in Proc. Conf. Neural Inf. Process. Mach. Learn., 2021, pp. 8359–8370.

[24] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, “Visual saliency transformer,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 4722–4732.

[25] Z. Wu, L. Su, and Q. Huang, “Cascade partial decoder for fast and accurate salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3902–3911.

[26] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, “Non-local deep features for salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6593–6601.

[27] B. Wang, Q. Chen, M. Zhou, Z. Zhang, X. Jin, and K. Gai, “Progressive feature polishing network for salient object detection,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 12128–12135.

[28] J. Zhang, J. Xie, and N. Barnes, “Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 349–366.

[29] A. Li, J. Zhang, Y. Lu, B. Liu, T. Zhang, and Y. Dai, “Uncertainty-aware joint salient object and camouflage object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 10066–10076.

[30] X. Tian, X. Zhao, H. Zhang, and H. Lu, “Multi-scale interactive network for salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 9413–9422.
[109] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, “Depth enhanced saliency detection method,” in Proc. Proc. Int. Conf. Multimedia Comput. Service, 2014, pp. 23–27.

[110] Y. Niu, Y. Geng, X. Li, and F. Liu, “Leveraging stereopsis for saliency analysis,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2012, pp. 454–461.

[111] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, “Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 5, pp. 2075–2089, May 2021.

[112] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, “Enhanced-alignment measure for binary foreground map evaluation,” in Proc. Int. Joint Conf. Artif. Intell., 2018, pp. 698–704.

[113] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, “Structure-measure: A new way to evaluate foreground maps,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 4548–4557.

[114] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 248–255.

Jing Zhang received the PhD degree in computer vision from Australian National University, in 2021. She is currently a lecturer with the School of Computing, Australian National University. Her main research interests include saliency detection, weakly supervised learning, generative models and uncertainty estimation. She won the Best Student Paper Prize at DICTA 2017, the Best Deep/Machine Learning Paper Prize at APSIPA ASC 2017 and the Best Paper Award Nominee at IEEE CVPR 2020.

Nick Barnes received the PhD degree in robotic vision from the University of Melbourne, in 1999. He is a professor with the School of Computing, Australian National University. He was a visiting researcher, LIRA Lab, University of Genoa, Italy in 1999; tenured lecturer with the University of Melbourne until 2003; with NICTA from 2003-16, he was a senior principal researcher and executive leader with the Computer Vision Research Group; and from 2016-19 led the Computer Vision Research Group, CSIRO. He has best paper awards and nominations from CVPR, RSS, IROS, and DICTA. With multiple patents in vision for prosthetic vision, he contributed the creation of Bionic Vision Technologies.

Jianwen Xie received the PhD degree in statistics from the University of California, Los Angeles, in 2016. He is a research manager with Akool Research, USA. He was a staff research scientist with Baidu Research, USA, from 2020 to 2022. Before joining Baidu, he was a senior research scientist with Hikvision Research, USA, from 2017 to 2020, and a postdoctoral researcher with the Center for Vision, Cognition, Learning, and Autonomy (VCLA), UCLA from 2016 to 2017. His research interests include focus on generative modeling and learning with applications in computer vision.

Ping Li received the master’s degree in computer science and the master’s degree in electrical engineering, and the two master’s degrees in earthquake geotechnical engineering and applied mathematics from the University of Washington, Seattle, WA, USA, and the PhD degree in statistics from Stanford University, in 2007. He was a recipient of the Young Instigator Award from the Office of Naval Research (ONR-YIP) and a recipient of the Young Investigator Award from the Air Force Office of Scientific Research (AFOSR-YIP).

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.