Generative Transformer for Accurate and Reliable Salient Object Detection

Yuxin Mao, Jing Zhang, Zhexiong Wan, Xinyu Tian, Aixuan Li, Yunqiu Lv, and Yuchao Dai, Member, IEEE

Abstract—We explore the impact of transformers on accurate and reliable salient object detection. For accuracy, we integrate the transformer with a deterministic model and delineate its advantages in structural modeling. Regarding reliability, we address the transformer’s tendency to produce overly confident, incorrect predictions. To gauge reliability implicitly, we introduce a latent variable model within the transformer framework, termed the inferential generative adversarial network (iGAN). The stochastic nature of the latent variable facilitates the estimation of predictive uncertainty, which serves as an auxiliary measure of the model’s prediction reliability. Different from the conventional GAN, which defines the distribution of the latent variable as fixed standard normal distribution $\mathcal{N}(0,1)$. The proposed “iGAN” infers the latent variable by gradient-based Markov Chain Monte Carlo (MCMC), namely Langevin dynamics, leading to an input-dependent latent variable model. We apply our proposed iGAN to fully supervised salient object detection, explaining that iGAN within the transformer framework leads to both accurate and reliable salient object detection. The source code and experimental results are publicly available via our project page: https://npucvr.github.io/TransformerSOD.

Index Terms—Vision transformer, salient object detection, inferential generative adversarial network.

I. INTRODUCTION

VISUAL salient object detection (SOD) [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16] aims to localize and segment regions of an image that attract human attention. It is typically framed as a binary segmentation task. Given the deterministic one-to-one mapping formulation, the main focus of conventional deep salient object detection models is achieving structure-preserving prediction with effective high/low-level feature aggregation. However, there are two major issues observed in existing salient object detection models: less effective global context modeling abilities and the over-confidence issue.

A. Less Effective Global Context Modeling Issue

Conventional CNN based saliency detection models usually consist of two main parts: 1) an encoder for feature extraction; and 2) a decoder for high/low feature aggregation, where the encoder is usually adopted from an ImageNet pre-trained backbone network, e.g., VGG [18], ResNet [17]. In this way, the SOD models are mainly designed to obtain effective decoders for feature aggregation [2], [3]. We visualize the different levels of CNN and transformer backbone features of the SOD models in Fig. 1 and find that the former encodes less accurate global context than the latter, especially for the large salient foreground (the first row of Fig. 1). The main reason for the better context modeling ability of the transformer lies in its self-attention mechanism, where global context is modeled. We find the better context modeling ability of the transformer is especially beneficial for context-based tasks, i.e. salient object detection.

B. Effective Context Modeling via Transformer

Researchers have found that the “Transformer” [19] has great potential to solve the limited receptive field issue in vision tasks. The advantage of the “Transformer” lies in the use of self-attention to capture global contextual information to establish a long-range dependency. Different from convolutional neural networks that focus on a small patch of the image, the transformer network [19] performs global context modeling with self-attention. Inspired by [20] and [21] and the accurate structure modeling ability of the transformer (see Fig. 1), we conduct extensive research to explore the contributions of the transformer for accurate salient object detection. Specifically, we design transformer based deterministic neural networks for SOD and explain that the accurate structure modeling and the global context modeling abilities lead to its superior performance (see Table II).

C. Overcoming the Over-Confidence Issue

Although significant performance has been achieved with the transformer, we still observe “over-confidence” issue within the transformer based SOD models, where the model tends to generate wrong predictions with high confidence, which is also defined as the model less-calibrated issue in [22].
Salient object detection: Driven by visual attention [31], UNet [33] structure, where effective decoders are designed pre-trained backbone networks [17], [18] as an encoder with related feature extraction. Deep SOD models usually take contrast [32]. Early works usually utilize this prior for saliency salient objects are defined as objects that have strong connection and adversarial attack and defense techniques. Specifically, we discuss FGSM [28], a gradient based attack, and perform adversarial training [29] to refine salient regions. Meanwhile, edge detection [46], [47] is likewise used as a piece of auxiliary information [41], [48] to improve the performance of SOD. Different attention mechanisms such as spatial and channel attention [49], [50], [51] or pixel-wise contextual attention [52] are also used to learn more discriminative features. Li et al. [53] propose dense attention to efficiently control the information propagation for SOD. DHNet [54] focuses on mining hard samples on salient object detection and proposes a dense sampling strategy for hard samples to construct a graph representation with samples from different classes and different confidence levels. Unlike the mainstream design refinement prediction networks, Zhang et al. [9] propose an automatic consolidation of multi-level features based on neural architecture search for flexible integration of information at different scales.

A. Vision Transformer and Its Applications

The transformer network [19] has sparked great interests in the computer vision community to adapt these models for vision tasks such as object detection [55], [56], [57], [58], [59], object tracking [60], [61], pose estimation [62], optical flow [63] etc. Inspired by the success of the Vision Transformer (ViT) [30] in image classification which splits the input image into a sequence of patches and feeds them to a standard Transformer encoder, some works extend transformers for dense prediction tasks, e.g., semantic segmentation or depth estimation. SETR [64] and PVT [58] use several convolutional layers as the decoder to upsample feature maps and get the dense prediction with the input image size. DPT [20] uses ViT [30] as an encoder to extract features from different spatial resolutions of the initial embedding. Liu et al. [21] present the Swin Transformer, a hierarchical transformer with a shifted windowing scheme to achieve an efficient network for vision tasks. Transformers are introduced in the SOD field to improve

We then present an inferential generative adversarial network (iGAN) to analyze the reliability degree [22] of the transformer based framework. Different from the conventional generative adversarial network (GAN) [23] which defines the distribution of the latent variable as fixed standard normal distribution \( \mathcal{N}(0, 1) \), our proposed “iGAN” infers the latent variable by gradient based Markov Chain Monte Carlo (MCMC), namely Langevin dynamics [24]. The latent variable within iGAN is sampled directly from its true posterior distribution [25], leading to more informative latent space exploration.

We apply the proposed iGAN to fully supervised SOD and explain that iGAN within the transformer framework leads to both accurate and reliable salient object detection, where the produced uncertainty maps [26] can serve as an auxiliary output to explain the reliability of model predictions. Experimental results show that the proposed iGAN within the transformer backbone can fix the “less effective global context modeling” and “over-confidence” issues, where the auxiliary uncertainty outputs can be used to explain model reliability. Further, to explain the robustness of the proposed generative saliency framework, we have conducted experiments on adversarial attack [27]. Specifically, we discuss FGSM [28], a gradient based attack, and perform adversarial training [29] to achieve model defense. Experimental results show robust performance of the proposed iGAN frameworks.

Our main contributions can be summarized as: 1) We extensively explore the contributions of transformer networks [19], [20], [21], [30] for accurate salient object detection and explain that the effective structure and global context modeling abilities lead to the superior performance of the transformer-based saliency detection network; 2) We present an inferential generative adversarial network (iGAN) to effectively measure the reliability degree of the transformer-based SOD network, leading to reliable saliency prediction; 3) We apply iGAN to fully supervised salient object detection to extensively explore the proposed new generative model within the transformer framework and test its robustness via using adversarial attack and defense techniques.

II. RELATED WORK

Salient object detection: Driven by visual attention [31], salient objects are defined as objects that have strong contrast [32]. Early works usually utilize this prior for saliency related feature extraction. Deep SOD models usually take the pre-trained backbone networks [17], [18] as an encoder with UNet [33] structure, where effective decoders are designed to achieve high-low level feature aggregation [2], [3], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44]. Among them, Wu et al. [2] propose a “stacked cross refinement network” and used the interaction between the edge module and the detection module to optimize the two tasks at the same time. Wei et al. [3] introduce an adaptive selection of complementary information when aggregating multi-scale features with a structure-aware loss function. Tang et al. [45] model the two tasks of discriminating salient regions and identifying accurate edges independently and solved the limitations of low-resolution SOD by using low-resolution images to delineate salient regions and using high-resolution to refine salient regions. Meanwhile, edge detection [46], [47] is likewise used as a piece of auxiliary information [41], [48] to improve the performance of SOD. Different attention mechanisms such as spatial and channel attention [49], [50], [51] or pixel-wise contextual attention [52] are also used to learn more discriminative features. Li et al. [53] propose dense attention to efficiently control the information propagation for SOD. DHNet [54] focuses on mining hard samples on salient object detection and proposes a dense sampling strategy for hard samples to construct a graph representation with samples from different classes and different confidence levels. Unlike the mainstream design refinement prediction networks, Zhang et al. [9] propose an automatic consolidation of multi-level features based on neural architecture search for flexible integration of information at different scales.
the performance [1], [65], [66], [67], [68]. Among them, [67], [68] focus on Co-SOD and RGB-D SOD. Despite previous explorations of transformers in salient object detection, our method uniquely integrates an inferential generative adversarial network (iGAN) and Langevin dynamics-based MCMC within the transformer framework, significantly enhancing accuracy and reliability.

B. Generative Models and Their Applications

There are two types of generative models, namely latent variable models [23], [69], [70] and energy-based models [71]. The former usually involves an extra latent variable to model the predictive distribution, and the latter directly estimates the compatibility of the input and output variable with a designed energy function. The variational auto-encoder (VAE) [69], [70] and generative adversarial network (GAN) [23] are two widely studied latent variable models. VAEs use an extra inference model to constrain the distribution of the latent variable, and GANs design a discriminator to distinguish the real samples and the generated samples. VAEs have already been successfully applied to image segmentation [72], [73] to produce stochastic predictions during testing. For saliency prediction, [74] adopts a VAE for image background reconstruction and the residual of the raw image and the reconstructed background is then defined as the salient region(s). Differently, [4] designs a conditional variational auto-encoder (CVAE) to model the subjective nature of saliency, where the latent variable is used to model the prediction variants. GAN-based methods can be divided into two categories, namely fully-supervised and semi-supervised settings. The former [75], [76] uses the discriminator to distinguish model predictions from ground truth, while the latter [77], [78] rely on the GAN to explore the contributions of unlabeled data. Reference [79] introduces an inferential Wasserstein GAN model, which is a principled framework to fuse auto-encoders and Wasserstein GAN and jointly learns an encoder network and a generator network motivated by the iterative primal-dual optimization process. Differently, we infer the latent variable via Langevin dynamics [24], which suffers no posterior collapse issue [80].

III. ACCURATE AND RELIABLE SALIENT OBJECT DETECTION VIA GENERATIVE TRANSFORMER

We define our training dataset as \( D = \{x_i, y_i\}_{i=1}^N \) of size \( N \), where \( x_i \) and \( y_i \) are the input RGB image and the corresponding ground truth saliency map, and \( i \) indexes the samples, which is omitted. Given any testing sample \( x^* \) with ground truth \( y^* \), we define the joint distribution of the model \( \theta \) as \( p(x^*, y^*, \theta|D) = p(y^*|x^*, \theta)p(\theta|D)p(x^*|D) \), where \( p(y^*|x^*, \theta) \) represents the predictive distribution or the inherent randomness given \( \theta \) as the oracle [26]. \( p(\theta|D) \) explains the ambiguity of the model \( \theta \) given the provided training dataset \( D \), and \( p(x^*|D) \) measures the discrepancy between \( x^* \) and the training dataset \( D \).

With the global context modeling ability of the transformer, the \( p(\theta|D) \) term can be modeled more effectively compared with the CNN frameworks. This improvement is attributed to the transformer’s global context awareness and its dynamic attention mechanism. However, there exists no solution in the transformer framework to model the predictive distribution \( p(y^*|x^*, \theta) \). Further, the training/testing discrepancy is not mentioned either, thus it is inconvenient to evaluate the domain gap caused by \( p(x^*|D) \). To fix the above-mentioned issues, we introduce a latent variable model with an extra latent variable \( z \) involved to model the inherent data noise. Specifically, with extra latent variable \( z \), the joint distribution of the testing sample \( x^* \) can be rewritten as:

\[
p(x^*, y^*, \theta, z|D) = \frac{p(x^*, y^*, \theta, z, D)}{p(D)} = \frac{p(y^*|x^*, \theta, z)p(x^*|\theta, z, D)p(\theta, z, D)}{p(D)} = \frac{p(y^*|x^*, \theta, z)p(\theta|D)p(z|x^*, D)p(x^*|D)}{p(D)}.
\]

The extra latent variable in Eq. (1) makes it convenient to estimate \( p(y^*|x^*, \theta, z) \), where the latent variable \( z \) can be sampled from \( p(z|x^*, D) \) during testing, modeling the discrepancy between training and testing sample.

In the following, we will first introduce a transformer for accurate saliency detection (Sec. III-A) for effective model parameter estimation (\( p(\theta|D) \)). We will introduce the latent variable model in Sec. III-B to achieve modeling of \( p(z|x^*, D) \), with which it’s convenient to model the inherent randomness of model prediction \( p(y^*|x^*, \theta, z) \). We present the objective function in Sec. III-D.

A. Transformer for Accurate Saliency Detection

The straightforward solution of using a transformer is to replace the CNN backbone with a transformer backbone, leading to the “transformer encoder”. We take the Swin transformer [21] as our transformer encoder, which takes the image as input and produces a list of feature maps \( f_{\theta_l}(x) = \{f^l_1, \ldots, f^l_4\} \) of channel size 128, 256, 512 and 1024 respectively, representing different levels of features. Different from [20], [30] that use fixed tokenization, the Swin transformer [21] is a hierarchical transformer structure whose representation is computed with self-attention in shifted non-overlapped windows thus it enables even larger receptive field modeling. Given the “transformer encoder”, we design a simple “convolution decoder” to achieve high/low-level feature aggregation. Specifically, we first feed each backbone feature \( f^l_1 \) to a simple convolutional block and obtain the new backbone feature \( \{f^l_1, \ldots, f^l_4\} \) of the same channel size \( C = 32 \). Such channel reduction operation aims to further enhance context modeling and reduce the huge memory requirement. Our final saliency map \( s = f_{\theta_o}(\{f^l_1, \ldots, f^l_4\}) \) is then obtained via a decoder parameterized by \( \theta_2 \).

The detailed structure of the decoder can be formulated as \( s = f_{\theta_{hid}}(f_{\theta_{cab}}(\{f^l_1, \ldots, f^l_4\})) \), where \( [\cdot] \) denotes the channel-wise concatenation operation, \( f_{\theta_{cab}} \) is the residual channel attention block [81], \( f_{\theta_{hid}} \) is the multi-scale dilated convolutional block [82] to obtain a one-channel saliency map. Note that, \( \theta = \{\theta_1, \theta_2\} \) indicates the entire parameters of our salient object detection network. \( f_{\theta_o}(x) \) can directly produce the saliency map for RGB image \( x \). With the global context
modeling ability, the regression ability of the transformer is proven better than CNN frameworks, leading to better \( \theta \) estimation given the same training dataset \( D \) with less ambiguity/uncertainty.

**B. Generative Model for Stochastic Saliency Detection**

As deep neural networks can fit any random noise [22], the deterministic CNN and transformer backbone based models have serious over-confidence issues, where the model could inaccurately assign a high probability to the wrong prediction. To overcome this issue, we desire a model that is aware of its prediction with reasonable predictive distribution modeling. As discussed in Eq. (1), a latent variable model makes it convenient to estimate predictive distribution \( p(y^*|x^*, \theta, z) \), which can be defined as being Gaussian distribution via:

\[
p(y^*|x^*, \theta, z) = \mathcal{N}(\mu(x^*, \theta, z), \sigma^2(x^*, \theta, z)),
\]

where the mean is \( \mu(x^*, \theta, z) = \mathbb{E}_{z \sim p(z|x^*, \theta)} p(y^*|x^*, \theta, z) \) and \( \sigma^2(x^*, \theta, z) = \mathbb{E}_{z \sim p(z|x^*, \theta)} (p(y^*|x^*, \theta, z) - \mu(x^*, \theta, z))^2 \) indicates the variance (uncertainty). Eq. (2) presents a convenient solution for evaluating uncertainty from a latent variable model, where the randomness or uncertainty of model prediction is controlled by the latent variable \( z \), making meaningful \( z \) quite desirable for reliable uncertainty estimation. We will first introduce the existing latent variable models and adapt them to our task. Then we analyze their advantages and limitations.

1) **Generative Adversarial Nets (GAN) [23]**: Within the GAN-based framework, we design an extra fully convolutional discriminator \( g_\beta \) following [78], where \( \beta \) is the parameter of the discriminator. Two different modules (the saliency generator \( f_\theta \) and the discriminator \( g_\beta \) in our case) play the minimax game in GAN based framework:

\[
\min_{f_\theta} \max_{g_\beta} V(g_\beta, f_\theta) = E_{(x,y) \sim \text{data}(x,y)}[\log g_\beta(y|x)] + E_{z \sim \text{p}(z)}[\log(1 - g_\beta(f_\theta(x, z)))],
\]

where \( \text{p}(z) \) is the joint distribution of training data, \( \text{p}(z) \) is the prior distribution of the latent variable \( z \), which is usually defined as \( \text{p}(z) = \mathcal{N}(0, I) \).

In practice, we define the loss function for the generator as the sum of a reconstruction loss \( L_{\text{rec}} \), and an adversarial loss \( L_{\text{adv}} \), which is \( L_{\text{gen}} = L_{\text{rec}} + \lambda L_{\text{adv}} \). The hyper-parameter \( \lambda \) is tuned, and empirically we set \( \lambda = 0.1 \) for stable training. For SOD, we define the reconstruction loss \( L_{\text{rec}} \) as the structure-aware loss as in Eq. (9), and the adversarial loss as a cross-entropy loss: \( L_{\text{adv}} = L_{\text{ce}}(g_\beta(f_\theta(x, z)), I) \), which fools the discriminator that the prediction is real, where \( I \) is an all-one matrix. The discriminator \( g_\beta \) is trained via the loss function: \( L_{\text{dis}} = L_{\text{ce}}(g_\beta(f_\theta(x, z)), 0) + L_{\text{ce}}(g_\beta(y), 1) \), which aims to correctly distinguish prediction and ground truth. Similarly, \( 0 \) is an all-zero matrix. In this way, the generator loss and the discriminator loss can be summarized as:

\[
L_{\text{gen}} = L_{\text{rec}} + \lambda L_{\text{adv}}, \quad L_{\text{dis}} = L_{\text{ce}}(g_\beta(f_\theta(x, z)), 0) + L_{\text{ce}}(g_\beta(y), 1).
\]

2) **Variational Auto-Encoder (VAE) [69]**: For dense prediction tasks with input variable \( x \) and output variable \( y \), the conditional variational auto-encoder (CVAE) [70], [83] treats the input image \( x \) as the conditional variable. As a conditional directed graph model, a conventional CVAE mainly contains two modules: a saliency generator model \( f_\theta(x) \) to produce the task related predictions, and an inference model \( q_\theta(z|x, y) \), which infers the latent variable \( z \) with image \( x \) and ground-truth \( y \) as input. Learning a CVAE framework involves approximation of the true posterior distribution of \( z \) with an inference model \( q_\theta(z|x, y) \), with the loss function as:

\[
L_{\text{cvae}} = \mathbb{E}_{z \sim q_\theta(z|x, y)}[-\log p_\theta(y|x, z)] - L_{\text{rec}} + D_{KL}(q_\theta(z|x, y) \parallel p_\theta(z|x)).
\]

The first term is the reconstruction loss and the second is the Kullback-Leibler divergence of prior distribution \( p_\theta(z|x) \) and posterior distribution \( q_\theta(z|x, y) \), where both of them are usually parameterized by multi-layer perceptron (MLP).

3) **Alternating Back-Propagation (ABP)**: Alternating back-propagation [25] updates the latent variable and network parameters in an EM manner. Given the network prediction with the current parameter set, it infers the latent variable by the Langevin dynamics based Markov Chain Monte Carlo (MCMC) [24], which is called “Inferential back-propagation”. Given the updated latent variable \( z \), the network parameter set is updated with gradient descent, which is called “Learning back-propagation” [25]. Similar to the VAE [69] or CVAE [70] frameworks, ABP intends to infer \( z \) and learn the network parameter \( \theta \) to minimize the reconstruction loss. Specifically, ABP [25] samples \( z \) directly from its posterior distribution with a gradient-based Monte Carlo method, namely Langevin Dynamics [24]:

\[
Z_{t+1} = Z_t + \frac{s_t^2}{2} \left[ \frac{\partial}{\partial Z} \log p_\theta(y, Z_t|x) \right] + s_t \mathcal{N}(0, I),
\]

where \( Z_0 \sim \mathcal{N}(0, I) \), and the gradient term is defined as:

\[
\frac{\partial}{\partial Z} \log p_\theta(y, Z|x) = \frac{1}{\sigma^2} (y - f_\theta(x, z)) \frac{\partial}{\partial z} f_\theta(x, z) - z.
\]

In the case of posterior collapse [80], this issue leads to the independence of the latent variable from the input image, resulting in a less representative latent space. In contrast, GAN-based models do not include an inference model. As a result, the latent variable \( z \) in GAN is always sampled from the standard normal distribution \( \mathcal{N}(0, I) \), which is less informative. This less informative latent space makes it less suitable for directly modeling the predictive distribution. However, the adversarial training strategy of GANs generally leads to better model performance compared to the other two
latent variable models. The ABP-based framework samples from the true posterior distribution using Eq. (6). However, the task-related training remains unchanged. Our experimental results indicate that this approach heavily influences deterministic performance, especially for conventional CNN backbone-based frameworks.

C. The Inferential GAN for Reliable Saliency Detection

In this paper, we introduce inferential generative adversarial network (iGAN), a new generative model for SOD, where we infer the latent variable within the proposed framework instead of defining it as fixed $\mathcal{N}(0, 1)$. Specifically, the proposed iGAN infers the latent variable by gradient-based Markov Chain Monte Carlo (MCMC), namely Langevin dynamics [24] (see Fig. 2), leading to an image conditioned latent variable. Further, we apply the adversarial training strategy in the training process, leading to reliable latent space exploration with fewer inference time steps.

Following the previous variable definitions, given the training example $(x, y)$, we intend to infer $z$ and learn the network parameters $\theta$ to minimize the reconstruction loss as well as a regularization term that corresponds to the prior on $z_t$. Our iGAN based framework includes three main parts: a generator for task related predictions, a discriminator to distinguish the prediction and ground truth, and an inference model via Langevin dynamics [24] to infer the latent variable with gradient based MCMC. Different from the isotropic Gaussian distribution assumption for the latent variable in GAN [23], or the possible posterior issue [80] within VAE [69], our latent variable $z$ is sampled directly from its real posterior distribution via gradient based MCMC following [25]. Further, we introduce extra adversarial loss and the fully convolutional discriminator, serving as a higher-order loss function for accurate deterministic predictions. Empirically, we set $s_t = 0.1$ and $\sigma^2 = 0.3$ in Eq. (6) and Eq. (7). During training, we sample $z_0$ from $\mathcal{N}(0, 1)$, and update $z$ via Eq. (6) by running $T = 5$ steps of Langevin sampling [24], and the final $z_T$ is then used to generate saliency prediction in our case. For testing, we can sample directly from the prior distribution $\mathcal{N}(0, 1)$.

1) Network Details: The proposed iGAN can be applied to any deterministic saliency detection model, and we show the flowchart of the proposed iGAN for saliency detection in Fig. 2. Specifically, we first extend the latent variable $z$ to the same spatial size as the highest level backbone feature ($t_4$ in this paper). Then, we concatenate $z$ with $t_4$ channel-wise and feed it to a $3 \times 3$ convolutional layer, which will serve as the new $t_4$ for saliency prediction. The discriminator contains four $3 \times 3$ convolutional layers following batch normalization and leakyReLU activation function with 64 channels, which takes the concatenation of image and model prediction (or ground truth) as input to estimate its pixel-wise realness. In this way, the discriminator loss in Eq. (4) can be rewritten as:

$$\mathcal{L}_{\text{dis}} = \mathcal{L}_{\text{ce}}(g_\beta([f_\theta(x, z), x]), 0) + \mathcal{L}_{\text{ce}}(g_\beta([y, x]), 1),$$

where $[\cdot, \cdot]$ is the channel-wise concatenation operation. The training of the proposed iGAN is the same as the conventional GAN based models in Eq. (4), except that we have an extra inference model via MCMC [24]. We show the learning pipeline of iGAN in Algorithm 1.

Algorithm 1 iGAN for Fully Supervised Saliency Detection

**Input:** (1) Training images $\{x_i\}_{i=1}^N$ with associated saliency maps $\{y_i\}_{i=1}^N$, where $i$ indexes images, and $N$ is the size of the training dataset. (2) Maximal number of learning epochs $E_P$. (3) Numbers of Langevin steps for posterior $T$; (4) Langevin step sizes for posterior $s_t$ and variance of inherent labeling noise $\sigma^2$.

**Output:** Parameters $\theta$ for the generator and $\beta$ for the discriminator.

1: Initialize $\theta$ and $\beta$
2: for $ep \leftarrow 1$ to $E_P$
3: Sample image-saliency pairs $\{(x_i, y_i)\}_{i=1}^N$
4: For each $(x_i, y_i)$, sample the prior $z_0^i \sim \mathcal{N}(0, 1)$, and sample the posterior $z_t^i$ using $T$ Langevin steps in Eq. (6) with a step size $s_t$ and inherent noise $\sigma^2$.
5: Update the transformer generator with model prediction $f_\theta(x_i, z_t^i)$ using the generator loss function in Eq. (4).
6: Update the discriminator with loss function in Eq. (8).
7: end for

2) Inferential GAN Analysis: Same as other generative models, iGAN aims to produce reliable uncertainty maps while keeping the deterministic performance unchanged. As the conventional GAN [23] has no inference step, the latent variable is independent of the input image $x$, leading to less informative uncertainty maps while sampling from the latent space at
test time. Although VAE [69] and ABP [25] can produce input-dependent latent space modeling, the possible posterior collapse [80] issue within the former and the less accurate deterministic prediction of the latter limit their applications for SOD. With the proposed iGAN, we can achieve two main benefits: 1) an extra inference step is included without increasing model parameters, leading to an input-dependent latent variable; 2) with the adversarial loss function serving as a high-order similarity measure, thus iGAN can lead to more effective model learning compared with ABP [25]. For the former, our iGAN is built upon GAN [23] and ABP [25], and the fully convolutional discriminator introduces less than 1M extra parameters, which is comparable to both the alternative latent variable models and the deterministic models. For the latter, the adversarial training is proven effective in maintaining the deterministic performance compared with the alternative stochastic models.

D. Objective Function

The objective is shown in Eq. (4), where the reconstruction loss \( L_{\text{rec}} \) is chosen as the structure-aware loss from [3], which is the sum of the weighted binary cross-entropy loss and the weighted IOU loss:

\[
L_{\text{rec}} = \omega(L_{\text{ce}}(s, y) + L_{\text{iou}}(s, y)),
\]

where \( y \) is the ground truth saliency map, \( \omega \) is the edge-aware weight, and is defined as \( \omega = 1 + 5\times(|ap(y) - y|) \), with \( ap(.) \) representing the average pooling operation. \( L_{\text{ce}} \) is the binary cross-entropy loss. \( L_{\text{iou}} \) is the weighted IOU loss [3]. The generator loss and discriminator loss are obtained following Eq. (4). Note that the latent variable \( z \) is updated via Langevin dynamics as shown in Eq. (6).

IV. EXPERIMENTS

A. Experimental Setting

1) Dataset: We train the models by using the DUTS training dataset [6] \( D = \{x_i, y_i\}_{i=1}^N \) of size \( N = 10,553 \), and test on six other widely used datasets: the DUTs testing dataset, ECSSD [84], DUT [85], HKU-IS [86], PASCAL-S [87] and the SOD testing dataset [88].

2) Evaluation Metrics: We use four evaluation metrics to measure the performance, including Mean Absolute Error \( M \), Mean F-measure \( F_p \), Mean E-measure \( E_k \) [94] and S-measure \( S_\alpha \) [95].

- **MAE** \( M \): is defined as the pixel-wise difference between the prediction \( s \) and the ground truth \( y \):
  
  \[ M = \frac{1}{HW} \sum_{c=0}^{W} |c - y|, \]

- **F-measure** \( F_p \): is a region-based similarity metric, and we provide the mean F-measure using varying fixed (0-255) thresholds.

- **E-measure** \( E_k \): is the recently proposed Enhanced alignment measure [94] in the binary map evaluation field to jointly capture image-level statistics and local pixel matching information.

- **S-measure** \( S_\alpha \): is a structure based measure [95], which combines the region-aware \( (S_r) \) and object-aware \( (S_o) \) structural similarity as their final structure metric: \( S_\alpha = \alpha S_o + (1 - \alpha) S_r \), where \( \alpha \in [0, 1] \) is set to 0.5 by default.

3) Implementation Details: We train our model in Pytorch with the Swin transformer backbone [21] pre-trained on ImageNet-1K [96], and other newly added layers are randomly initialized. We resize all the images and ground truth to 384 × 384. The maximum epoch is 50 and the training mini-batch size is set to 6. The initial learning rates are 2.5 × 10⁻⁵ with AdamW [97] optimizer.

B. Accurate and Reliable Salient Object Detection

1) Performance Comparison With Benchmark Models: We compare the proposed framework with benchmark salient object detection models in Table I. Note that, VST [65] and GT SOD [66] are two existing transformer based saliency detection models. We observe the competitive performance of our CNN based generative model (CIGAN) with existing techniques in Table I. To focus on explaining the superior performance of the transformer backbone for SOD, our decoder has only 1M parameters, which is around 5% of model parameters of existing techniques. Further, as can be seen from Table I, our TIGAN achieves optimal and suboptimal results across most metrics and datasets, leading to a better performance of our generative model (TIGAN) compared with other transformer based SOD methods [1], [65], [93]. Different from the deterministic methods [1], [65], [93], as a generative model, we aim to produce stochastic predictions leading to reliable saliency prediction. In this way, we compare with GT SOD [66], another generative transformer SOD model, in the way of both accurate and reliable saliency prediction. Table I shows that the proposed iGAN achieves comparable performance compared with GT SOD [66], leading to an alternative generative saliency transformer. Both our method and GT SOD belong to generative saliency solutions, where the predictions are stochastic, and different sampling steps can lead to relatively different performance comparisons. We can focus on that the \( S_{\text{alpha}} \) values in Table I of our proposed model are higher than GT SOD on most datasets (except HKU-IS). In addition, in Fig. 3, we further visualize the produced uncertainty maps of GT SOD [66] and ours for SOD. The more reliable uncertainty maps, highlighting the less confident or hard regions, further explain our superiority. Besides the visual comparison, we also explain the robustness of our proposed model via adversarial attack and defense and show the results in Fig. 8, which further explains the superiority of the generative transformer.

2) Accurate Saliency Model: In Sec. I, we discuss that the CNN backbone is not effective in detecting salient objects that rely on global context and the stride and pooling operation lead to less accurate structure information of CNN backbone features. We then compare the performance of CNN backbone (B_cnn with ResNet50 [17] backbone) and transformer backbone (B_tr with Swin transformer backbone [21]) for RGB image based SOD, and show performance in Table II, where the models share the same decoder.¹ Note that, for both B_cnn and B_tr, we use binary cross-entropy loss for the

¹We adjust the decoder accordingly to the backbone features.
TABLE I

| Method         | DUTS [6] | ECSSD [84] | DUT [85] | HKU-IS [86] | PASCAL-S [87] | SOD [88] |
|----------------|----------|------------|----------|-------------|---------------|----------|
| SCRNet [3]     | 0.85     | 0.93       | 0.90     | 0.90        | 0.93          | 0.90     |
| FSNet          | 0.88     | 0.92       | 0.95     | 0.95        | 0.94          | 0.95     |
| ITSSD [9]      | 0.86     | 0.91       | 0.90     | 0.90        | 0.94          | 0.97     |
| CPD [12]       | 0.85     | 0.90       | 0.91     | 0.91        | 0.93          | 0.91     |
| AFNet [50]     | 0.86     | 0.87       | 0.91     | 0.91        | 0.93          | 0.91     |
| EGNNet [88]    | 0.88     | 0.86       | 0.91     | 0.91        | 0.94          | 0.91     |
| MINet [102]    | 0.84     | 0.83       | 0.91     | 0.91        | 0.93          | 0.91     |
| LDF [91]       | 0.89     | 0.86       | 0.92     | 0.92        | 0.93          | 0.92     |
| GateNet [38]   | 0.85     | 0.90       | 0.95     | 0.95        | 0.96          | 0.95     |
| MSNet [9]      | 0.87     | 0.92       | 0.95     | 0.95        | 0.96          | 0.95     |
| CTDNet [92]    | 0.89     | 0.92       | 0.95     | 0.95        | 0.96          | 0.95     |
| PAKIN [10]     | 0.90     | 0.93       | 0.95     | 0.95        | 0.96          | 0.95     |
| CIGAN          | 0.87     | 0.87       | 0.93     | 0.93        | 0.95          | 0.93     |
| VST [65]       | 0.86     | 0.90       | 0.91     | 0.91        | 0.92          | 0.91     |
| SRformer [93]  | 0.91     | 0.89       | 0.91     | 0.91        | 0.94          | 0.91     |
| ICON [1]       | 0.89     | 0.87       | 0.91     | 0.91        | 0.93          | 0.91     |
| GTSSOD [66]    | 0.90     | 0.87       | 0.91     | 0.91        | 0.93          | 0.91     |
| TIGAN          | 0.91     | 0.87       | 0.91     | 0.91        | 0.93          | 0.91     |

Fig. 3. Performance comparison with existing generative SOD model GTSOD [66], where prediction in each block is the model prediction and the corresponding uncertainty map.

Fig. 4. Predictions of CNN and transformer backbone models without (B_cnn and B_tr) and with (B'_cnn and B'_tr) structure-aware loss.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE III
RELIABLE FULLY-SUPERVISED RGB SOD MODELS, WHERE WE PRESENT PERFORMANCE OF STOCHASTIC SALIENCY PREDICTION MODELS VIA GAN, CVAE, ABP AS WELL AS THE PROPOSED iGAN. PERFORMANCE OF THE BASELINE DETERMINISTIC MODELS (B’_cnn and B’_tr in Table II) ARE LISTED FOR EASIER REFERENCE

| Method | DUTS [6] | ECSSD [84] | DUT [85] | HKU-IS [96] | PASCAL-S [67] | SOD [83] |
|--------|----------|------------|----------|-------------|---------------|----------|
| B’_cnn | S_4, F_4, F_2, E_1 | S_4, F_4, F_2, E_1 | S_4, F_4, F_2, E_1 | S_4, F_4, F_2, E_1 | S_4, F_4, F_2, E_1 | S_4, F_4, F_2, E_1 |
| CGAN  | 0.82, 0.84, 0.91, 0.03 | 0.92, 0.91, 0.94, 0.03 | 0.82, 0.74, 0.53, 0.05 | 0.92, 0.91, 0.94, 0.03 | 0.82, 0.85, 0.89, 0.06 | 0.82, 0.85, 0.86, 0.03 |
| CCVAE | 0.87, 0.93, 0.11, 0.04 | 0.92, 0.92, 0.94, 0.04 | 0.87, 0.73, 0.54, 0.06 | 0.92, 0.92, 0.94, 0.04 | 0.87, 0.84, 0.89, 0.06 | 0.87, 0.84, 0.86, 0.03 |
| CABB | 0.82, 0.75, 0.85, 0.05 | 0.88, 0.77, 0.93, 0.05 | 0.78, 0.67, 0.51, 0.07 | 0.87, 0.85, 0.93, 0.07 | 0.81, 0.78, 0.84, 0.09 | 0.77, 0.74, 0.79, 0.10 |
| CIGAN | 0.87, 0.82, 0.96, 0.04 | 0.92, 0.91, 0.94, 0.03 | 0.82, 0.73, 0.54, 0.06 | 0.91, 0.89, 0.93, 0.04 | 0.86, 0.86, 0.89, 0.06 | 0.83, 0.81, 0.86, 0.07 |

For easier reference, we also include the baseline models B’_cnn and B’_tr from Table II in Table III, and the CNN and transformer backbone based stochastic models are built upon the two baseline models respectively. Table III shows that the four types of generative models can achieve comparable deterministic performance (compared with the corresponding deterministic baseline models) for SOD. As the goal of a generative model is to obtain stochastic predictions for the model explanation, the deterministic performance of the proposed iGAN may be slight. The main reason lies in two parts. First, the hyper-parameters within the inference model in Eq. (6) need to be tuned to effectively explore the latent space. Second, the final performances of those generative models are obtained by performing multiple iterations (10 iterations in this paper) of forward passes during testing, and the performance

distributed compactly around the image center, leading to less significant performance gain with the transformer backbone.

3) Reliable Saliency Model: Model reliability is an important factor for measuring accountability for decisions before deployment, and a reliable model should be aware of its predictions. In this paper, we introduce the iGAN for reliable saliency detection with an image conditioned latent prior. In addition to the proposed iGAN, we also design GAN-based [23], CVAE-based [69], [70] and ABP-based [25] generative models for SOD. The performance is shown in Table III respectively, where “CGAN”, “CCVAE”, “CABB” and “CIGAN” are the stochastic models based on GAN, CVAE, ABP, and the proposed inferential GAN respectively with CNN backbone, and “TGAN”, “TCVAE”, “TABP” and “TIGAN” are the transformer counterparts.
Fig. 6. Predictions from different generative models in Table III, where we randomly sample $T = 10$ times and obtain the entropy of mean prediction as predictive uncertainty [26]. Note that the predictions within each method block are saliency prediction and uncertainty, respectively.

### TABLE IV

| Model Analysis Related experiments, Where We Discuss Model Performance With Respect to Model Optimizer ("CIGAN_SGD" and "TIGAN_SGD"), Initialization Weights ("CIGAN_R", TIGAN_R and TIGAN_22K) and Different Transformer Backbones ("TIGAN_ViT") |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Method                          | DUTS [6]        | ECSSD [84]      | DUT [85]        | HKU-IS [86]     | PASCAL-S [87]   | SOD [88]        |
| CIGAN                           | S_{a} \uparrow F_{a} \uparrow E_{a} \uparrow M \downarrow | S_{a} \uparrow F_{a} \uparrow E_{a} \uparrow M \downarrow | S_{a} \uparrow F_{a} \uparrow E_{a} \uparrow M \downarrow | S_{a} \uparrow F_{a} \uparrow E_{a} \uparrow M \downarrow | S_{a} \uparrow F_{a} \uparrow E_{a} \uparrow M \downarrow | S_{a} \uparrow F_{a} \uparrow E_{a} \uparrow M \downarrow |
| CIGAN_SGD                       | 876.82          | 906.04          | 923.913.945.037 | 823.732.488.061 | 911.892.943.034 | 836.836.895.068 |
| CIGAN_R                         | 876.82          | 913.947.038     | 822.731.844.066 | 905.891.944.036 | 853.837.898.070 | 826.810.865.079 |
| TIGAN                           | 748.62          | 775.108         | 835.792.851.097 | 736.603.764.112 | 822.757.861.082 | 754.695.779.138 |
| TIGAN_SGD                       | 912.873.941.026 | 941.936.967.025 | 861.796.890.047 | 929.922.964.023 | 879.869.919.053 | 861.854.894.066 |
| TIGAN_ViT                       | 897.859.938.033 | 926.921.952.034 | 852.784.888.058 | 919.907.954.031 | 865.854.907.063 | 831.816.864.077 |
| TIGAN_22K                       | 770.869.802.099 | 850.821.872.084 | 752.635.786.107 | 841.801.886.072 | 762.717.796.136 | 728.687.765.144 |

of the mean prediction is then reported, which varies with different iterations of sampling.

We argue that the main advantages of generative models lie in their stochastic nature, making it possible to estimate predictive uncertainty [26] for model reliability estimation. In Fig. 6, we visualize the uncertainty maps of each generative model. In this paper, the “uncertainty” refers to predictive uncertainty [26], [98], which is the total uncertainty, including both data uncertainty and model uncertainty. Given the mean predictions after multiple forward passes during testing, the predictive uncertainty is defined as the entropy of the mean prediction. A reliable model should be aware of its prediction, leading to a reasonable uncertainty model to explain model prediction. Fig. 6 shows that the uncertainty map from the proposed inferential GAN explains better model prediction, highlighting the hard samples caused by training/testing discrepancy. Especially, for the 1st sample, although closely touched, the top and bottom instances have different saliency degrees, which is relatively different from the training dataset. However, most of the closely touched objects in the training dataset share a similar saliency degree. The ground truth of the 2nd image is biased, focusing only on a compact foreground region, where the less compact region is discarded. All four latent variable models can discover the discarded less compact region, where the uncertainty map of “TIGAN” is more informative in explaining the less accurate predictions.

### C. Discussions

In this section, we discuss some training variants based on our proposed CIGAN and TIGAN, including model optimizers, weight initializations, and transformer backbones. The results are shown in Table IV.

1) Model Performance w.r.t. Optimizer: We observe that the AdamW optimizer is more suitable to train the transformer backbone (Swin [21] in particular) based framework compared with SGD. To explain this, we train CIGAN and TIGAN with SGD optimizer, leading to CIGAN_SGD and CIGAN_SGD respectively in Table IV. We observe that for both the CNN and transformer backbone based networks, the SGD optimizer usually achieves worse performance compared with the AdamW optimizer. Note that models with the two types of optimizers share the same initial learning rate, and the momentum rate in this paper is set to 0.9. We find that after the first epoch, the AdamW optimizer based model jumps directly to a minimum of smaller loss compared with SGD, and later, the loss decreases behaviors of both models are similar. We also tried different learning rate configurations for models with the two types of optimizers, and the performance of SGD based model is still bad. We further explore whether the AdamW converges faster than SGD. However, even when we train more epochs for SGD based model, the conclusion is still similar. We will investigate it further to extensively explain the different model behaviors with various types of optimizers.

2) The Importance of Initialization Weights: For both CNN and transformer backbone, we initialize the weights with the image classification model trained on the ImageNet-1K [96] dataset. To test how the initialization weights contribute to the model performance, we randomly initialize the two models (CIGAN and TIGAN in Table II) and obtain model...
performance as CIGAN_R and TIGAN_R in Table IV. We observe the worse performance of both CIGAN_R and TIGAN_R, which further illustrates the necessity of fine-tuning the backbone models for SOD. We also initialize our transformer backbone with parameters pre-trained on the ImageNet-22K dataset and show the result as TIGAN_22K in Table IV. The better performance of TIGAN_22K compared with TIGAN again explains the importance of the initialization weights.

3) Different Transformer Backbones Analysis: Following our pipeline, we change the Swin transformer backbone [21] to the ViT backbone [20], [30], and achieve TIGAN_ViT in Table IV. Note that, the ViT backbone we used in Table IV is initialized with weights trained on ImageNet-22K. The comparable performance of TIGAN_ViT compared with TIGAN_22K explains that the two types of backbones both work for SOD. However, since the SOD task usually uses the model initialized by ImageNet-1K as the backbone, for the sake of fair comparison, we do not put the results of these two models into the comparison of the main model.

4) Model Performance w.r.t. Training Datasets Scales: As the two types of backbones have significantly different numbers of model parameters, leading to different model capacities. We aim to analyze how model capacity is sensitive to the scales of the training dataset. We then train our transformer backbone with parameters pre-trained on the ImageNet-22K dataset and show the result as TIGAN_22K in Table IV. The better performance of TIGAN_22K compared with TIGAN again explains the importance of the initialization weights.

5) Robustness to Adversarial Attack: Deep neural network models are known to suffer from adversarial examples. With small perturbations, model predictions can be changed drastically [27]. Common defense methods for adversarial attacks include adversarial training [29], certified robustness [99], etc. In this paper, we investigate model robustness with respect to adversarial attacks. Specifically, we discuss FGSM [28], a gradient based attack, and perform adversarial training [29] to achieve model defense.

FGSM [28] attack only needs to do backpropagate once to get the gradient of classification loss with respect to the input $x$, and the adversarial sample $x_{adv}$ can be generated via:

$$x_{adv} = x + \varepsilon \text{sign}(\nabla_x \mathcal{L}(\theta, x, y)),$$

where $\mathcal{L}(\theta, x, y)$ is the classification loss, and the sign function is used to achieve faster convergence. Correspondingly, the adversarial training based defense is achieved via training the model with adversarial sample pair $(x_{adv}, y)$, leading to a new objective:

$$\min_{\theta} \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[ \max_{\delta \in S} \mathcal{L}(\theta, x + \delta, y) \right],$$

where $S$ is the candidate adversarial attacks. In practice, a more efficient way to achieve adversarial training based defense is thorough joint training with both clean sample $x$ and adversarial sample $x_{adv}$ with weighted loss:

$$\min_{\theta} \mathbb{E}_{(x, y) \sim \mathcal{D}} \left[ \alpha \mathcal{L}(\theta, x, y) + (1 - \alpha) \mathcal{L}(\theta, x_{adv}, y) \right],$$

where $\alpha = 0.5$ is used to control the balance of accurate prediction ($\mathcal{L}(\theta, x, y)$) and model robustness ($\mathcal{L}(\theta, x_{adv}, y)$). This equal weighting ensures that the training process gives importance to both maintaining accuracy on normal inputs and enhancing robustness against adversarial attacks.

In this paper, we perform adversarial attack with FGSM [28] and defense with adversarial training in Eq. (12) to both the baseline models (B'_cnn and B'_tr in Table III) and the proposed iGAN based frameworks (CIGAN and TIGAN in Table III). Specifically, based on the above models, we set $\varepsilon = 8/255$ in Eq. (10) following [28] to generate adversarial samples $x_{adv}$ of each training image. The generated adversarial samples will be used to train the above models again together with the clean sample $x$ to achieve the defense process. We report model performance in Table V. Note that the performance of attacked models is obtained by performing FGSM [28] attack on the testing samples, where the adversarial testing samples are fed to the specific model to generate the predictions. We show in Fig. 8 the clean sample $x$ and its prediction $s$, the adversarial sample $x_{adv}$ and its prediction $s_{adv}$,
TABLE V

| Method          | DUTS [6] | ECSSD [84] | DUT [85] | HKU-IS [86] | PASCAL-S [87] | SOD [88] |
|-----------------|----------|------------|----------|-------------|---------------|---------|
|                 | $S_0 \uparrow F_0 \uparrow E_0 \uparrow M \downarrow$ | $S_0 \uparrow F_0 \uparrow E_0 \uparrow M \downarrow$ | $S_0 \uparrow F_0 \uparrow E_0 \uparrow M \downarrow$ | $S_0 \uparrow F_0 \uparrow E_0 \uparrow M \downarrow$ | $S_0 \uparrow F_0 \uparrow E_0 \uparrow M \downarrow$ | $S_0 \uparrow F_0 \uparrow E_0 \uparrow M \downarrow$ |
| B'$_{cnn}$      | 882.840 | 916.097 | 922.919 | 947.055 | 823.742 | 851.057 | 912.901 | 947.073 | 855.841 | 896.065 | 832.825 | 863.073 |
| CIGAN           | 878.782 | 906.042 | 925.913 | 945.037 | 823.733 | 848.061 | 911.892 | 943.034 | 856.836 | 895.068 | 833.816 | 862.075 |
| B'$_{tr}$       | 913.882 | 947.026 | 939.940 | 965.024 | 860.301 | 894.045 | 922.922 | 964.023 | 876.972 | 917.053 | 855.853 | 897.059 |
| TIGAN           | 912.372 | 941.026 | 941.936 | 967.025 | 861.786 | 890.047 | 929.922 | 964.023 | 879.869 | 919.053 | 844.854 | 894.060 |

Fig. 8. Model robustness to adversarial attack, i.e. FGSM [28] attack, where the samples (from left to right) within each method (CIGAN and TIGAN) are the prediction of $x$, the adversarial sample $x_{adv}$, its prediction $s_{adv}$, and the prediction of $x_{adv}$ after adversarial training.

Fig. 9. Failure cases of the transformer backbone compared with the CNN backbone (B'$_{cnn}$ and B'$_{tr}$), and iGAN compared with CGAN within the CNN backbone (CIGAN and CGAN) for RGB salient object detection.

6) Failure Case Analysis: To further investigate the limitations of both the transformer backbone and the iGAN framework, we look deeper in Table III, and the predictions of each related model. We find two main issues: 1) the transformer backbone does not always perform superior to the CNN backbone; 2) our iGAN model can lead to over-smoothed predictions compared with CGAN due to the diverse generation process.

For the former, to exclude the influence of the iGAN framework, we compare the predictions of the CNN backbone (B'$_{cnn}$) and transformer backbone (B'$_{tr}$) for SOD, and show samples in Fig. 9. We observe more false positives within B'$_{tr}$, which can be explained as the “double-edged sword” of the transformer backbone. On the one hand, the larger receptive field of the transformer makes it superior in localizing the larger salient foreground. On the other hand, the less salient objects that expand for a larger region can be falsly detected as positive foreground. We argue that salient object ranking [100] can be beneficial in identifying the less salient regions by providing extra saliency degree evaluation. For the latter, as the latent variable $z$ within iGAN is conditioned on input $x$, leading to informative latent space, where the predictive distribution $p(y^*|x^*, \theta, z)$ for input $x^*$ has larger variance compared with the CGAN based framework. In this case, the averaged prediction is over-smoothed. Similarly, this issue can also be fixed with saliency ranking [100].

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed an inferential GAN within the transformer framework for fully supervised salient object detection. Different from typical GANs that define the prior distribution as a standard normal distribution, we inferred the latent variable via Langevin Dynamics [24], a gradient based MCMC, leading to the image-conditioned prior distribution. Through extensive experiments, we observed that a larger receptive field of the transformer leads to a better performance on images with larger salient objects (see Fig. 5). However, we also found the double-edged sword effect of the larger receptive field that leads to serious false positives (see Fig. 9). Extensive experimental results demonstrate the superiority of our transformer backbone-based generative network, achieving new benchmarks with reliable uncertainty maps.

Our proposed generative model aims to estimate the reliability of salient object prediction with uncertainty maps, which also show superiority in achieving robust models (see Table V, Fig. 8 and Fig. 8). Our future work includes two main parts. Firstly, we will apply the produced uncertainty map (see Fig. 6) to the saliency generator for effective

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
hard negative mining. In this way, the uncertainty map can not only explain model predictions but also serve as an important prior for effective model learning. Secondly, we have several hyper-parameters within the inference model, and we observe that our model performance can be influenced by them. We plan to further investigate model performance w.r.t. those hyper-parameters. Our proposed iGAN model can also be extended to other tasks, such as Multi-Human Parsing tasks [101,102,103].

ACKNOWLEDGMENT

The authors would like to thank Dr. Dengping Fan for the discussion and help during the manuscript submission process.

REFERENCES

[1] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, “Salient object detection via integral learning,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 3, pp. 3738–3752, Mar. 2023.
[2] Z. Wu, L. Su, and Q. Huang, “Stacked cross refinement network for edge-aware salient object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 7264–7273.
[3] J. Wei, S. Wang, and Q. Huang, “F²Net: Fusion, feedback and focus for salient object detection,” in Proc. AAAI Conf. Artif. Intell., 2020, vol. 34, no. 7, pp. 12321–12328.
[4] J. Zhang et al., “Uncertainty inspired RGB-D saliency detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 9, pp. 5761–5779, Sep. 2022.
[5] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, “Weakly-supervised salient object detection via scribble annotations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 12543–12552.
[6] C. Wang et al., “Learning to detect salient objects with image-level supervision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3796–3805.
[7] J. Zhang, T. Zhang, Y. Daf, M. Harandi, and R. Hartley, “Deep unsupervised saliency detection: A multiple noisy labeling perspective,” in IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 9029–9038.
[8] D. T. Nguyen et al., “DeepUSPS: Deep robust unsupervised saliency prediction with self-supervision,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 1–11.
[9] M. Zhang, T. Liu, Y. Piao, S. Yao, and H. Lu, “Auto-MSFNet: Search multi-scale fusion network for salient object detection,” in Proc. 29th ACM Int. Conf. Multimedia, Oct. 2021, pp. 667–676.
[10] B. Xu, H. Liang, R. Liang, and P. Chen, “Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection,” in Proc. AAAI Conf. Artif. Intell., 2021, vol. 35, no. 4, pp. 3004–3012.
[11] Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Chen, and B. Ren, “EDN: Salient object detection via extremely-downsampled network,” IEEE Trans. Image Process., vol. 31, pp. 3125–3136, 2022.
[12] Z. Wu, L. Su, and Q. Huang, “Cascaded partial decoder for fast and accurate salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3902–3911.
[13] Z. Wang, Y. Zhang, Y. Liu, D. Zhu, S. A. Coleman, and D. Kerr, “ELWNet: An extremely lightweight approach for real-time salient object detection,” IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 11, pp. 6404–6417, Nov. 2023.
[14] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, “Uncertainty-aware joint salient object and camouflage object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 10066–10076.
[15] A. Li, Y. Mao, J. Zhang, and Y. Dai, “Mutual information regularization for weakly-supervised RGB-D salient object detection,” IEEE Trans. Circuits Syst. Video Technol., vol. 34, no. 1, pp. 397–410, Jan. 2023.
[16] X. Tian, J. Zhang, M. Xiang, and Y. Dai, “Modeling the distributional uncertainty for salient object detection models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 19660–19670, Jun. 2023.
[93] Y. K. Yan and W. Lin, “Towards a complete and detail-preserved salient object detection,” *IEEE Trans. Multimedia*, vol. 26, pp. 4667–4680, 2024.

[94] 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. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.

[95] 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. (ICCV)*, Oct. 2017, pp. 4548–4557.

[96] 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.*, Jun. 2009, pp. 248–255.

[97] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv:1711.05101*, 2017.

[98] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft, “Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning,” in *Proc. ICLR*, 2018, pp. 1184–1193.

[99] E. Wong and Z. Kolter, “Provable defenses against adversarial examples via the convex outer adversarial polytope,” in *Proc. ICLR*, 2018, pp. 5286–5295.

[100] M. A. Islam, M. Kalash, and N. D. B. Bruce, “Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7142–7150.

[101] J. Li et al., “Multi-human parsing with a graph-based generative adversarial network,” *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 1, pp. 1–21, Feb. 2021.

[102] J. Li et al., “Multi-human parsing machines,” in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 45–53.

[103] J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, “Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing,” in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 792–800.

---

**Yuxin Mao** received the Bachelor of Engineering degree from Southwest Jiaotong University in 2020. He is currently pursuing the Ph.D. degree with the School of Electronics and Information, Northwestern Polytechnical University, Xi’an, China. He received the Best Paper Award Nominee at ICIUS 2019.

**Jing Zhang** is currently a Lecturer with the School of Computing, Australian National University, Canberra, Australia. Her main research interests include saliency detection, weakly supervised learning, and generative models. She received the Best Paper Award Nominee at IEEE CVPR 2020, the Best Student Paper Prize at DICTA 2017, and the Best Deep/Machine Learning Paper Prize at APSIPA ASC 2017.

**Zhexiong Wan** received the Bachelor of Engineering degree from Northwestern Polytechnical University, Xi’an, China, in 2019, where he is currently pursuing the Ph.D. degree with the School of Electronics and Information.

**Yunqiu Lv** received the B.E. degree in statistics and the M.S. degree in computer science and technology from Xidian University, Xi’an. She is currently pursuing the Ph.D. degree in signal and information processing with the School of Electronics and Information, Northwestern Polytechnical University, Xi’an. Her current research interests include machine learning and computer vision.

**Xinyu Tian** received the B.E. degree in electronic information engineering from Northwestern Polytechnical University in 2021, where she is currently pursuing the master’s degree in signal and information processing with the School of Electronics and Information. Her main research interests include uncertainty modeling and object detection.

**Aixuan Li** received the M.E. degree in signal and information processing from Northwestern Polytechnical University in 2022, where she is currently pursuing the Ph.D. degree majoring in signal and information processing with the School of Electronics and Information. Her main research interests include salient/camouflaged object detection and adversarial attack.

**Yuchao Dai** (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in signal and information processing from Northwestern Polytechnical University (NPU), Xi’an, China, in 2005, 2008, and 2012, respectively. He was an ARC DECRA Fellow with the Research School of Engineering, Australian National University, Canberra, Australia. He is currently a Professor with the School of Electronics and Information, NPU. His research interests include structure from motion, multi-view geometry, low-level computer vision, deep learning, compressive sensing, and optimization. He received the Best Paper Award in IEEE CVPR 2012, the Best Algorithm Prize in NRSFM Challenge at CVPR 2017, 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. He served as the Area Chair for CVPR, ICCV, ECCV, and NeurIPS.