SuperVAE: Superpixelwise Variational Autoencoder for Salient Object Detection

Bo Li, Zhengxing Sun,*[*] Yuqi Guo
State Key Laboratory for Novel Software Technology, Nanjing University, China

Abstract

Image saliency detection has recently witnessed rapid progress due to deep neural networks. However, there still exist many important problems in the existing deep learning based methods. Pixel-wise convolutional neural network (CNN) methods suffer from blurry boundaries due to the convolutional and pooling operations. While region-based deep learning methods lack spatial consistency since they deal with each region independently. In this paper, we propose a novel salient object detection framework using a superpixelwise variational autoencoder (SuperVAE) network. We first use VAE to model the image background and then separate salient objects from the background through the reconstruction residuals. To better capture semantic and spatial contexts information, we also propose a perceptual loss to take advantage from deep pre-trained CNNs to train our SuperVAE network. Without the supervision of mask-level annotated data, our method generates high quality saliency results which can better preserve object boundaries and maintain the spatial consistency. Extensive experiments on five wildly-used benchmark datasets show that the proposed method achieves superior or competitive performance compared to other algorithms including the very recent state-of-the-art supervised methods.

Introduction

As a fundamental but challenging problem, salient object detection is derived with the goal of discovering and locating distinctive objects or regions in an image which attract human attention. It endows many high-level computer vision systems with the capability to take advantage of human attention for more promising processing and analysis, such as object recognition (Ren et al. 2014), image semantic segmentation (Wei et al. 2017), visual tracking (Hong et al. 2015) and video summarization (Mademlis, Tefas, and Pitas 2017), etc.

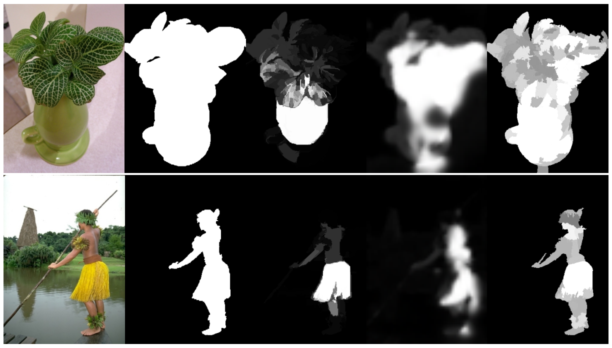


Figure 1: (a) Input images, (b) Ground truth masks, (c-e) Results of (c) MDF (Li and Yu 2016b), (d) RFCN (Wang et al. 2018), and (e) our method.

Conventional saliency detection methods usually utilize hand-crafted features such as color, texture, contrast to represent the visual properties of pixels or regions. These methods then distinguish salient objects from background according to the extracted features with heuristic priors and cues. Despite their great success, however, it is rather difficult for these methods to capture the semantic and structural information of salient objects in images because of the hand-crafted low-level features they use. When encountering complex images, such as images with low-contrast objects or cluttered backgrounds, these methods usually produce unsatisfying saliency results.

Recently, deep learning led a revolution in computer vision for its superior performance in object recognition and classification (Russakovsky et al. 2015). By incorporating deep networks, such as CNN, or fully convolutional neural network (FCN) into salient object detection algorithms (Li and Yu 2016a; Wang et al. 2018), with the extracted semantic features the detection accuracy has been improved rapidly in comparison with previous state-of-the-art results. However, while pixel-wise saliency prediction networks are good to evaluate objectness in an image, they lack the capability to generate clear boundary for salient object. As shown in Figure 1(d), after the stridden convolution and pooling operations, these methods lose much signifi-

[*] Corresponding author is Zhengxing Sun(szx@nju.edu.cn). This work was supported by National High Technology Researchand Development Program of China (No.2007AA01Z334), National Natural Science Foundation of China (Nos.61321491 and 61272219), Innovation Fund of State Key Laboratory for Novel Software Technology (Nos. ZZKT2013A12, ZZKT2016A11 and ZZKT2018A09)

cant location information and many fine details of objects, leading to coarse and blurry boundary results. This problem can be alleviated in some region-based methods like superpixel-wise classification networks (Li and Yu 2016b; Zhao et al. 2015). However, independently processing each superpixel makes these methods lack of ability to capture the spatial contexts information of superpixels. Therefore, it may cause failure detection results in some complex scenes. And the problem is also illustrated in Figure 1(c). Moreover, to ensure the performance, both pixel-wise and region-based networks rely on a great quantity of mask-level annotations. It is expensive, time-consuming and laborious for a fundamental visual task like salient object detection.

In this paper, a novel salient object detection framework is proposed, in which, a superpixel-wise variational autoencoder network is used to address these aforementioned challenges. Specifically, following the basic rule of photographic composition that the image boundary is mostly background, we design a superpixel-wise VAE to model image background. Then, the salient object detection can be formulated as an estimation of reconstruction residuals of all superpixels in an image using VAE network. More importantly, inspired by the recent neural style transfer works (Gatys, Ecker, and Bethge 2016), we propose a perceptual loss in training our SuperVAE network to take advantage from deep pre-trained CNNs to capture more semantic and spatial contexts information for superpixels. Unlike previous region-based networks which deal with each superpixel of an image independently, the perceptual loss of each superpixel in our SuperVAE network is calculated within the entire image. In this manner, our methods can maintain the spatial consistency of superpixels and generate high quality salient object detection results with clear boundaries. Meanwhile, our framework gets rid of the severe reliance on mask-level annotated data.

The main contributions of this work can be summarized as follows.

- We introduce a novel salient object detection framework. By adopting a superpixel-wise variational autoencoder network, we can separate salient objects from image background effectively.
- We propose a perceptual loss to capture more semantic and spatial contexts information from deep pre-trained CNNs for training our SuperVAE network.
- The proposed method generates high quality saliency results which can better preserve object boundaries and maintain the spatial consistency. Extensive experiments on five widely-used benchmark datasets show that our method achieves superior or competitive performance compared to other algorithms including the very recent state-of-the-art supervised methods.

Related Work

In this section, representative works including the recent deep learning methods in salient object detection are reviewed. Traditional methods treat salient object detection as a low level vision problem. Most of them are based on low-level manually designed features, such as color, gradient, texture, contrast, etc. Then, heuristic priors such as s the center prior (Liu et al. 2011) and the background prior (Wei et al. 2012) (Wei et al. 2012) or high-level knowledge like objectness (Jia and Han 2013) are incorporated with hand-crafted low-level features to predict good salient scores. Despite the efficient saliency cues these methods used, the representation ability of low-level feature on semantic information limits their performance in complex scenes, such as cluttered backgrounds and low-contrast imaging patterns. More comprehensive analysis of these low-level feature based traditional methods are summarized in a survey paper (Borji et al. 2015).

Recent years, deep learning has attracted a lot of attention for its outstanding performance in computer vision tasks. Based on the supervision of thousands of pixel-level saliency map annotations, the deep learning based saliency methods can easily capture the semantic information and produce more accurate saliency prediction results. For example, inspired by the great success of fully convolutional networks (FCNs) (Long, Shelhamer, and Darrell 2015), Wang et al (2018) developed a recurrent fully convolutional network to predict saliency maps based on the original image and then stage-wisely refine the prediction results of the last recurrent step. Zhang et al (Zhang et al. 2017b) proposed a bidirectional framework with a novel dropout technique to learn the deep uncertain convolutional features, which can improve the robustness and accuracy of saliency detection. However, while pixel-wise dense saliency prediction is efficient, the resulting saliency maps are coarse and with blurry object boundaries. It is because the features they used at deep layers of CNN lose location and fine details information of objects due to the multiple stridden convolution and pooling operations. Region-based methods alleviate this problem by better maintaining the location information through pre-segmentation. Li et al (Li and Yu 2016b) proposed multi-scale deep features by extracting features of all superpixels at three scales and then fused them to generate their saliency scores. Lee et al (Lee, Tai, and Kim 2016) used hand-crafted low-level features to form a low level distance map. They concatenated the encoded low level distance map and the high level features to complement the low-level detail information, and then used a fully connected neural network classifier to evaluate the saliency of a query superpixel. However, these methods simply process each superpixel in an independent way, which cause the loss of spatial contexts information. Lacking spatial consistency usually leads to detection errors when the salient object consists of several different parts. Moreover, the added hand-crafted low-level features are too subjective to capture the spatial consistency in all the images.

Our method well addresses these challenges in the aforementioned deep learning based methods. Not only our superpixel-wise salient object detection framework well preserves the location information, but also the proposed perceptual loss can extract more effective detail and semantic information from different layers in the pre-training network. Besides, we calculate the perceptual loss of superpixels within the entire image and can better maintain the

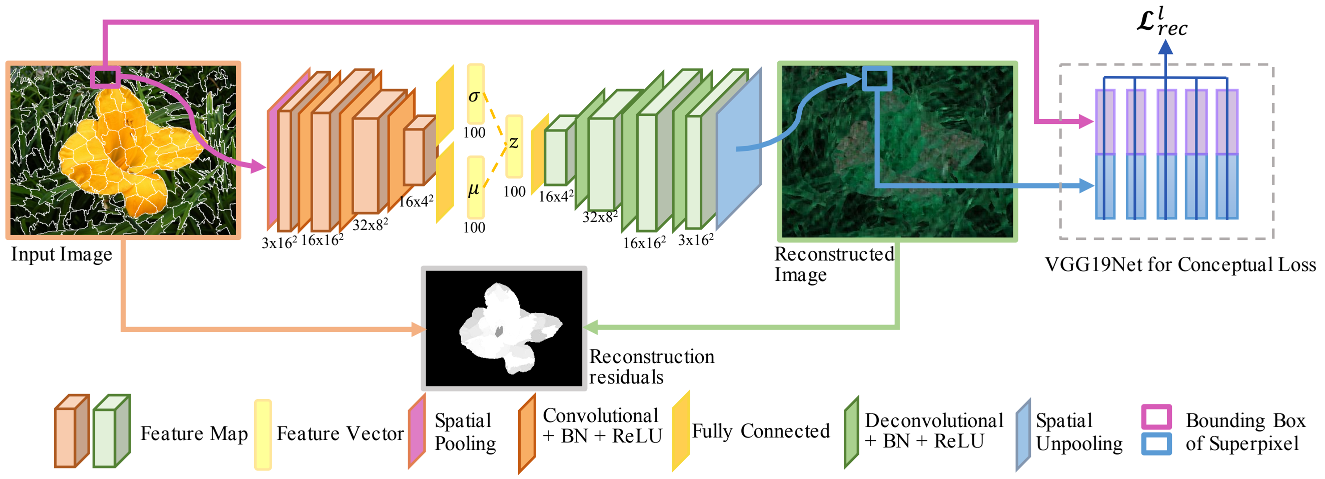


Figure 2: The main framework of our proposed superpixel-wise variational autoencoder network.

spatial consistency. In the meantime, unlike aforementioned deep learning based methods, the proposed framework also gets rid of excessive dependence on mask-level annotated data.

Method

The overview of our method is illustrated in Figure 2. For an input image, we first over segment it into superpixels by SLIC algorithm (Achanta et al. 2012). We then select the background samples from the image boundary. These background superpixels are used to train the proposed VAE network through the perceptual loss. Next, we use the trained VAE to reconstruct all superpixels in the input image and calculate the reconstruction residual. To handle the scale problem of images in different sizes, we compute the reconstruction residuals at different scales and fuse them to generate the final saliency results. In the following subsections, we will elaborate our superpixel-wise VAE network with the perceptual loss and how to generate the final saliency results through the deep reconstruction residuals.

Variational Autoencoder

Essentially, the true aim of salient object detection is to find objects that are distinctive from the image background (Han et al. 2015). So, it is natural to come up with the idea of modeling the property of background first and thereby separating salient objects from the background. Following this idea, some methods (Han et al. 2015; Lu et al. 2016) intuitively consider image saliency detection as an estimation of reconstruction error of whole image with the learned background model. However, because of lacking generalization ability, their background models such as sparse coding and denoising autoencoders may fail to well reconstruct some background regions which are variations of learned background. This leads to inaccurate salient scores in some regions. To handle this problem, we propose to use Variational Autoencoder (VAE) (Kingma and Welling 2013) in our framework to model image background. As a probabilistic generative neural network, VAE can give calibrated probabilities, while models of previous methods are deterministic discriminative model that do not have a probabilistic foundation. Specifically, VAE encoder network maps an input sample x to a distribution over latent variables $z \sim Enc(x) = q(z|x)$. And decoder network maps from this latent space distribution to the original input space $\bar{x} \sim Dec(z) = p(x|z)$. Both $q(z|x)$ and $p(x|z)$ are commonly assumed to be Gaussian distribution. Therefore, instead of predicting a single latent vector z , VAE predicts two vectors μ and σ and sample $z = \mu + \sigma \odot \varphi$, where φ is standard Gaussian (zero mean, unit variance) and $\odot$ is element-wise multiplication. Thus, compared with the previous models, VAE can model the background patterns efficiently and generate well variations of background.

To train the VAE model, first we need to maximize the marginal log-likelihood of each observation in x , and the VAE reconstruction loss $\mathcal{L}_{rec}$ is the negative expected log-likelihood of the observations in x. Meanwhile, the difference between the distribution of $q(z|x)$ and the distribution of a Gaussian distribution $p(z)$, which is called Kullback-Leibler divergence, needs to be minimized to control the distribution of the latent vector z. Therefore, VAE models can be trained by optimizing the sum of the reconstruction loss $\mathcal{L}_{rec}$ and KL divergence loss $\mathcal{L}_{kl}$.

$$\mathcal{L}_{rec} = -\mathbb{E}_{q(z|x)} \log p(x|z) \tag{1}$$

$$\mathcal{L}_{kl} = D_{KL}(q(z|x)||p(z)) \tag{2}$$

$$\mathcal{L}_{vae} = \mathcal{L}_{rec} + \mathcal{L}_{kl} \tag{3}$$

Superpixel-Wise VAE Network Architecture

As we discussed before, region-based methods can better capture the location information and preserve object boundary, compared with the pixel-wise CNN methods. So we design a superpixel-wise VAE network with a symmetric CNN architecture. To be specific, we construct 3 convolutional layers in the encoder network with 3 × 3 kernels. The stride of first layer is 1, while the strides of other two layers are set to be 2 to achieve spatial downsampling instead of using deterministic spatial functions such as pooling. Each convolutional layer is followed by an adaptive batch normalization layer and a ReLU activation layer. Then, to compute the KL divergence loss and sample latent variable z, we add two fully-connected output layers (for mean and variance) to encode. As a symmetric architecture, we use 3 deconvolutional layers with the same 3×3 kernel size in the decoder network. For upsampling we use nearest neighbor method by a scale of 2 instead of standard zero-padding. We also add batch normalization layer and ReLU activation layer following deconvolutional layer to help stabilize training. The details of our superpixel-wise VAE architecture is shown in Figure 2.

However, an image superpixel may have an irregular shape with variable pixel number inside, while the convolutional layer usually requires fixed-length inputting vector. To obtain available inputting for each superpixel, we first generate the bounding box of a superpixel. Moreover, to make the inputting vector only relevant to the pixels inside the superpixel, we fill the pixels outside the region but still inside its bounding box with the mean pixel values of superpixel. Then, we further perform average spatial pooling over an adaptive grid as with (He et al. 2015). We divide the bounding box of a superpixel into $h \times w$ cells. Let the size of the bounding box be $H \times W$. Average spatial pooling is performed within each cell with $H/h \times W/w$ pixels. Afterwards, the aggregated feature vector of each superpixel has $h \times w \times 3$ dimensions.

Perceptual Loss

As defined in neural style transfer (Gatys, Ecker, and Bethge 2016), perceptual loss of two images is the difference between the hidden features in a pre-trained deep convolutional neural network Φ. The core idea of feature perceptual loss is to seek the consistency between the hidden representations of two images. In our method, we propose to use the perceptual loss to capture important features such as detail and semantic information from the hidden representations in different layers. Thus, our superpixel-wise VAE network can better model the background and reconstruct high quality superpixel results. Then we can separate salient objects from the background more accurately through the reconstruction residuals. Specifically, the 19-layer VGGNet (Simonyan and Zisserman 2014) is chosen as loss network Φ to construct perceptual loss, which is trained for classification problem on ImageNet dataset. Let x represents a superpixel in image I, and $\bar{x}$ represents corresponding VAE output. Contrary to our previous process on x, we first perform spatial unpooling over $\bar{x}$ to restore it to its original superpixel size. Then we fill this reconstructed superpixel $\bar{x}$ into its corresponding location in the original image I and obtain a new image I'. Next, let $\Phi(I)^l$ represents the feature map of the l^{th} hidden layer when input image I is fed to network Φ. The perceptual loss for one layer ($\mathcal{L}^l_{rec}$) between superpixel x and $\bar{x}$ can be calculated through the different(squared euclidean distance) of $\Phi(I)^l$ and $\Phi(I')^l$,

$$\mathcal{L}^l_{rec} = \frac{1}{RF^l_x \times C^l}(\Phi(I)^l - \Phi(I')^l)^2 \; , \tag{4}$$

where C^l is the number of channels of the feature map $\Phi(I)^l$ and RF^l_x represents the receptive field size of superpixel x in feature map $\Phi(I)^l$. To accelerate the calculation of perceptual loss, we can fill more than one reconstructed superpixels into image I as long as their receptive fields in l^{th} hidden layer do not overlap with each other. Then the only thing we need to do is replacing RF^l_x with the sum of receptive fields of all filled superpixels in I'. As can be seen, unlike previous region-based networks which simply process each superpixel in an independent way, we compute the perceptual loss of each superpixel within the entire image. By this mean, our method can capture the spatial contexts information to better maintain the spatial consistency.

The final reconstruction loss is defined as the total loss by combining different layers of VGG Network. We also add a pixel-by-pixel reconstruction loss(cross entropy) between superpixel x and $\bar{x}$ in RGB space to capture low-level detail information. For a uniformly expression, we use $\mathcal{L}^0_{rec}$ to represent this loss. To train our superpixel-wise VAE, we jointly minimize the KL divergence loss $\mathcal{L}_{kl}$ and the reconstruction loss $\mathcal{L}^l_{rec}$ for different layers,

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{kl} + \sum_i^l (w_l \mathcal{L}^l_{rec}) \; , \tag{5}$$

where α and w_l are weighting parameters for KL Divergence and image reconstruction loss in different layers.

Saliency Detection via Reconstruction Residual

As described, after modeling the image background through our superpixel-wise VAE network, we reconstruct all superpixels in the input image and then separate salient objects from the background through the reconstruction residuals. Just like what we do in training network, we estimate the difference between superpixel x and reconstruction result $\bar{x}$ not only in RGB space but also in hidden representations as the reconstruction residuals. Let $r^l(x)$ represents the reconstruction residual (squared euclidean distance) of superpixel x calculated in l^{th} hidden layer. Then the final reconstruction residual is defined as the combination of reconstruction residuals in different hidden representations,

$$r(x) = \sum_i^l (w_l r^l(x)) \; , \tag{6}$$

where w_l are the weighting parameters for reconstruction residuals in different layers, same as the weighting parameters in calculation of image reconstruction loss. After normalization, the salient scores ε_x for each superpixel in the input image can be obtained.

To handle the scale problem of images in different sizes, inspired by Lu et al. (2016), we compute the reconstruction residuals at different scales and fuse them to generate the final saliency results. For a full-resolution saliency map, we assign saliency to each pixel by integrating results from multi-scale superpixel-level saliency

$$E(z) = \frac{\sum_{s=1}^{N_s} \omega_{zn^{(s)}} \varepsilon_{n^{(s)}}}{\sum_{s=1}^{N_s} \omega_{zn^{(s)}}}, \omega_{zn^{(s)}} = \frac{1}{\|f_z - v_{n^{(s)}}\|_2}, \quad (7)$$

where N_s is the number of scale, and $\varepsilon_{n^{(s)}}$ is the superpixel-level saliency in scale s. f_z is the low-level detail features of pixel z, which consist of Lab and RGB color features and coordinate of z. $n^{(s)}$ denotes the label of the superpixel containing pixel z at scale s and $v_{n^{(s)}}$ is the mean low-level detail features of all pixels in superpixel $n^{(s)}$. Thus, $\omega_{zn^{(s)}}$ regards the similarity of pixel z with its corresponding superpixel as the weight to average the reconstruction residuals in multi-scale.

Experiments

Implementation

Our proposed SuperVAE network has been implemented on the basis of pytorch, an open source framework for CNN training and testing. When we select superpixels on the image boundary as background samples to train our VAE network, we calculate the boundary connectivity (Zhu et al. 2014) for each superpixel and then use it to eliminate incorrect background examples on image boundary. For multi-scale reconstruction residuals, we generate superpixels at three different scales (150, 250, 350 superpixels). All superpixels are resized to $16 \times 16 \times 3$ by average spatial pooling. For reconstruction loss, we use the combination of $\mathcal{L}_{rec}^0$ and perceptual loss from layers $relu3\_1$, $relu4\_1$, $relu5\_4$, which are called $\mathcal{L}_{rec}^3$, $\mathcal{L}_{rec}^4$, $\mathcal{L}_{rec}^5$ respectively. The weighing parameters w_0, w_3, w_4, w_5 of reconstruction loss are set to 0.2, 0.1, 0.2, 0.5 respectively in our experiments. To ensure our SuperVAE model can learn as much information from reconstruction loss as it can, we adopt a KL cost annealing method (Bowman et al. 2016) in our experiments. Specifically, we add a variable weight to the KL term in the loss function at training time. At the start of training, we set that weight to zero, and then, as training progresses, we gradually increase this weight by 0.01 until it reaches 1.

We run our method on an octa-core PC machine with an NVIDIA GTX 1080Ti GPU and an i7-6900 CPU. During the training, we use ADAM stochastic gradient optimization method with batch size 10, and learning rate 0.005. For a single input image, the training process usually converges in 200 iterations. The average training time for each image is 6.2s, then it takes 1.3s seconds for the trained model to detect salient objects in the input image with 400x300 pixels.

Datasets and Evaluation Criteria

We evaluate the performance of our method on five public datasets: **ECSSD** (Shi et al. 2016) dataset contains 1,000 natural images, in which many semantically meaningful and complex structures are included. **ASD** (Achanta et al. 2009) consists of 1000 images and most images in this dataset contains single object. **SED** (Borji et al. 2015) dataset has two non-overlapped subsets, i.e., SED1 and SED2. SED1 has 100 images each containing only one salient object, while SED2 has 100 images each containing two salient objects. **SOD** (Wang et al. 2017) dataset contains 300 images and it was originally designed for image segmentation. Many images in this dataset have multiple salient objects with low contrast.

To evaluate the performance of varied methods, we adopt three metrics, including the widely used precision-recall (PR) curves, F-measure and mean absolute error (MAE) (Borji et al. 2015). The PR curve of a specific dataset exhibits the mean precision and recall of saliency maps at different thresholds. The F- measure is a weighted mean of average precision and average recall, calculated by

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}. \quad (8)$$

We set β^2 to be 0.3 as suggested in (Borji et al. 2015).

For fair comparison on non-salient regions, we also calculate the mean absolute error (MAE) by

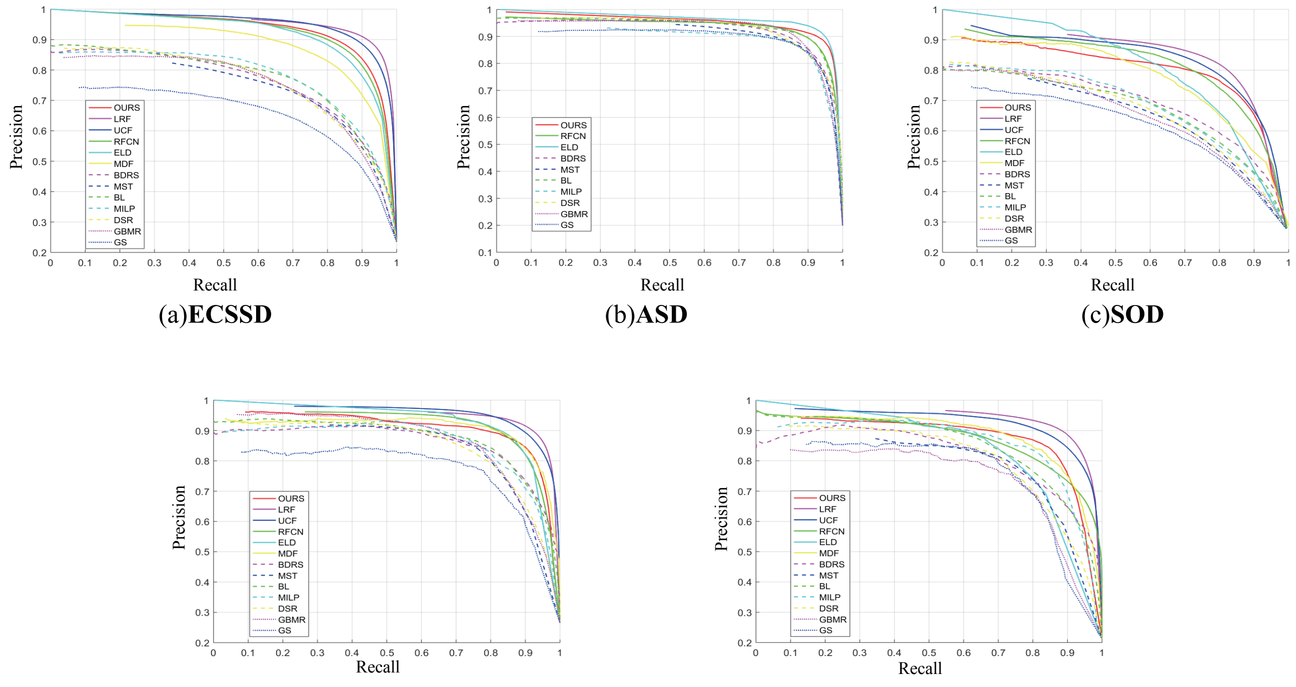
$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)|, \quad (9)$$

where W and H are the width and height of the input image. $S(x,y)$ and $G(x,y)$ are the pixel values of the saliency map and the binary ground truth at (x,y), respectively.

Comparison with the State-of-the-arts

To fully evaluate the detection performance, we compare our proposed method with other 12 state-of-the-art ones, including 6 deep learning based algorithms (**LRF** (Zhang et al. 2018), **UCF** (Zhang et al. 2017b), **RFCN** (Wang et al. 2018), **ELD** (Lee, Tai, and Kim 2016), **MDF** (Li and Yu 2016b), **BDRS** (Han et al. 2015)) and 6 conventional algorithms (**MST** (Tu et al. 2016), **BL** (Lu et al. 2017), **MILP** (Huang et al. 2017), **GBMR** (Zhang et al. 2017a), **DSR** (Lu et al. 2016), **GS** (Wei et al. 2012)). For fair comparison, we use either the implementations with recommended parameter settings or the saliency maps provided by the authors.

Quantitative Evaluation. In order to better demonstrate the characteristics of our work, we divide state-of-the-art methods into two groups. Group 1 contains one deep learning based method BDRS and all conventional methods which don't need pixel-level supervision. While Group 2 consists of 5 deep learning based methods including LRF, UCF , RFCN , ELD , MDF which need the supervision of pixel-level saliency map annotations. As part of the quantitative evaluation, we first evaluate our method using precision-recall curves. As shown in Fig. 3, our method significantly outperforms all unsupervised methods (dashed line) on all datasets and obtains a competitive ranking among supervised methods. Moreover, a quantitative comparison of maximum F-measure and MAE is listed in Table. 1. Our proposed SuperVAE method improves the F-measure achieved by the best-performing existing unsupervised algorithm by 17.5%, 1.03%, 1.73%, 2.73% and 15.8% respectively on ECSSD, ASD, SED1, SED2 and SOD. And at the same time, it lowers the MAE by 34.39%, 11.69%, 6.92%, 19.2% and 28.6% respectively on ECSSD, ASD, SED1, SED2



(a)**ECSSD** (b)**ASD** (c)**SOD**

(d)**SED1** (e)**SED2**

Figure 3: The PR curves of the proposed algorithm and other state-of-the-art methods.

Supervision	Methods	ECSSD		ASD		SED1		SED2		SOD	
		F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
Unsupervised	OURS	**0.828**	**0.103**	**0.867**	**0.068**	**0.822**	**0.121**	**0.789**	**0.101**	**0.733**	**0.159**
	MST (Tu et al. 2016)	0.705	0.157	0.859	0.098	0.808	0.130	0.768	0.125	0.633	0.223
	BDRS (Han et al. 2015)	0.549	0.281	0.682	0.175	0.659	0.202	0.622	0.197	0.517	0.301
	BL (Lu et al. 2017)	0.605	0.216	0.752	0.129	0.780	0.185	0.661	0.181	0.547	0.267
	MILP (Huang et al. 2017)	0.651	0.177	0.844	0.077	0.741	0.152	0.753	0.129	0.560	0.243
	DSR (Lu et al. 2016)	0.644	0.171	0.828	0.080	0.731	0.158	0.715	0.140	0.551	0.234
	GBMR (Zhang et al. 2017a)	0.640	0.190	0.837	0.085	0.759	0.166	0.707	0.169	0.549	0.239
	GS (Wei et al. 2012)	0.553	0.206	0.745	0.109	0.660	0.176	0.696	0.150	0.527	0.251
Pixel-level Annotations	LRF (Zhang et al. 2018)	0.880	0.054	–	–	0.902	0.051	0.871	0.052	0.789	0.124
	UCF (Zhang et al. 2017b)	0.844	0.078	–	–	0.865	0.065	0.810	0.069	0.738	0.148
	RFCN (Wang et al. 2018)	0.824	0.107	0.863	0.070	0.850	0.117	0.767	0.113	0.743	0.170
	ELD (Lee, Tai, and Kim 2016)	0.810	0.080	0.886	0.037	0.871	0.153	0.759	0.104	0.712	0.164
	MDF (Li and Yu 2016b)	0.807	0.105	–	–	0.779	0.123	0.768	0.115	0.721	0.192

Table 1: Quantitative comparison with the state-of-the-arts on five famous benchmark datasets. The bold and underlined numbers indicate the best and the second best results, respectively. "–" means corresponding methods are trained on that dataset.

and SOD. When compared with the supervised methods in group 2, our method is weaker than LRF and UCF. But it is noteworthy that as an unsupervised framework we consistently outperform MDF and achieve competitive performance comparing with ELD and RFCN .

Qualitative Evaluation. We further provide some typical saliency maps of different methods to intuitively demonstrate advantages of the proposed SuperVAE over other methods, as shown in Figure 4. From these results, we can observe that our method is more effective to detect the salient objects accurately, and obtain more clear object boundaries for input images. Specifically, most of the compared methods can not predict the whole objects in the first two rows images, while our method captures the whole salient regions by better maintaining the spatial consistency. For the images in the 3-4 rows, most of the compared methods wrongly assign high salient scores to background regions, while our method better suppresses the salient score of background regions and preserve clear object boundaries. The image in the last row is challenging with complicated background and a low-contrast salient object, and our method still accurately detect the salient object with the effective information in the hidden representations.

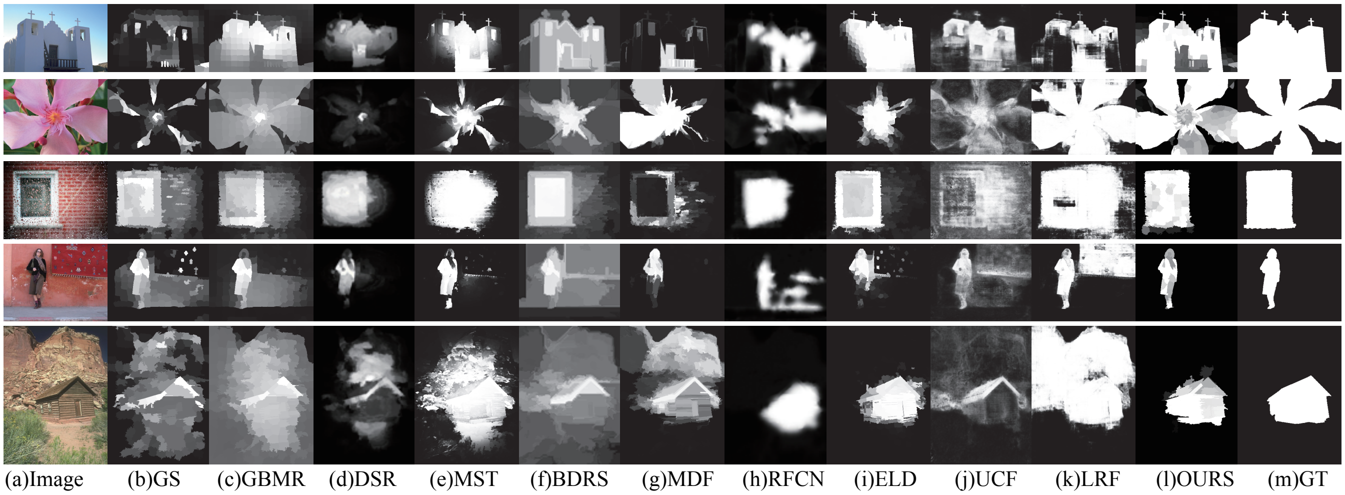


Figure 4: Comparison of typical saliency maps. (a) original image, (m) ground truth mask. Due to the limitation of space, we don't show the results of BL and MILP.

Models	ECSSD		ASD		SED1		SED2		SOD	
	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
OURS-BS	0.813	0.127	0.862	0.078	0.811	0.129	0.781	0.116	0.715	0.191
OURS-MF	0.821	0.109	0.865	0.069	0.820	0.123	0.785	0.104	0.730	0.164
OURS-PL	0.759	0.138	0.861	0.072	0.813	0.128	0.776	0.122	0.707	0.196
OURS	0.828	0.103	0.867	0.068	0.822	0.121	0.789	0.101	0.733	0.159

Table 2: The F-measure and MAE of different settings on five salient object detection datasets.

Ablation Analysis

To verify the effectiveness of the key components in proposed model, we perform ablation experiments on the datasets mentioned above. Specifically, we compared our final results(OURS) with the results without background samples selection (OURS-BS), the results without multi-scale fusion (OURS-MF), and the results without using perceptual loss (OURS-PL). In **OURS-BS**, we train our VAE network directly using the superpixels on the image boundary, without using the boundary connectivity to eliminate incorrect examples. In **OURS-MF**, we generate the saliency result for an input image in three different scales individually, and the results are not fused. The quantitative evaluation results of OURS-MF are calculated by averaging the quantitative evaluation results in three scales. In **OUR-PL**, we only use the reconstruction loss in RGB space ($\mathcal{L}^0_{rec}$) to train our SuperVAE network without the perceptual loss in hidden representations and the reconstruction residuals are also calculated only in RGB space.

The experimental results are shown in Table. 2, from which we can see that: (1) OUR-BS obtains worse performance than our final results, which demonstrates the effectiveness of background samples selection. (2) Without multi-scale fusion, the average F-measure decrease about 0.036 and MAE increase about 0.04, respectively, which means our multi-scale strategy can well handle the scale problem of images in different sizes. (3) Without using perceptual loss, the performance (in terms of the F-measure and MAE) of the proposed framework gets worse more obviously, demonstrating the significance of the semantic and spatial contexts information in the hidden representations. Meanwhile, even only using $\mathcal{L}^0_{rec}$ loss, OUR-PL still outperforms all unsupervised methods in group 1, which demonstrates the effectiveness of our SuperVAE network.

Conclusions

In this paper, we propose a novel salient object detection framework using SuperVAE network. Our method first use SuperVAE to model the image background and then separate salient objects from the background through the reconstruction residuals. For training, we also propose a perceptual loss to take advantage of deep pre-trained CNNs to better capture semantic and spatial contexts information. As a result, our method generates high quality saliency maps which can better preserve object boundaries and maintain the spatial consistency, without the supervision of mask-level annotated data. Extensive experiments on five famous benchmark datasets show that the proposed method achieves superior or competitive performance compared to other algorithms including the very recent state-of-the-art supervised methods.

References

Achanta, R.; Hemami, S. S.; Estrada, F. J.; and Süsstrunk, S. 2009. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604.

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE TPAMI* 34(11):2274–2282.

Borji, A.; Cheng, M.; Jiang, H.; and Li, J. 2015. Salient object detection: A benchmark. *IEEE TIP* 24(12):5706–5722.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A.; Jozefowicz, R.; and Bengio, S. 2016. Generating sentences from a continuous space. In *SIGNLL*, pages 10–21.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423.

Han, J.; Zhang, D.; Hu, X.; Guo, L.; Ren, J.; and Wu, F. 2015. Background prior-based salient object detection via deep reconstruction residual. *IEEE TCSV* 25(8):1309–1321.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE TPAMI* 37(9):1904–1916.

Hong, S.; You, T.; Kwak, S.; and Han, B. 2015. Online tracking by learning discriminative saliency map with convolutional neural network. In *ICML*, pages 597–606.

Huang, F.; Qi, J.; Lu, H.; Zhang, L.; and Ruan, X. 2017. Salient object detection via multiple instance learning. *IEEE TIP* 26(4):1911–1922.

Jia, Y., and Han, M. 2013. Category-independent object-level saliency detection. In *ICCV*, pages 1761–1768.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *CoRR* abs/1312.6114.

Lee, G.; Tai, Y.; and Kim, J. 2016. Deep saliency with encoded low level distance map and high level features. In *CVPR*, pages 660–668.

Li, G., and Yu, Y. 2016a. Deep contrast learning for salient object detection. In *CVPR*, pages 478–487.

Li, G., and Yu, Y. 2016b. Visual saliency detection based on multiscale deep CNN features. *IEEE TIP* 25(11):5012–5024.

Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; and Shum, H. 2011. Learning to detect a salient object. *IEEE TPAMI* 33(2):353–367.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440.

Lu, H.; Li, X.; Zhang, L.; Ruan, X.; and Yang, M. 2016. Dense and sparse reconstruction error based saliency descriptor. *IEEE TIP* 25(4):1592–1603.

Lu, H.; Zhang, X.; Qi, J.; Tong, N.; Ruan, X.; and Yang, M. 2017. Co-bootstrapping saliency. *IEEE TIP* 26(1):414–425.

Mademlis, I.; Tefas, A.; and Pitas, I. 2017. Summarization of human activity videos using a salient dictionary. In *ICIP*, pages 625–629.

Ren, Z.; Gao, S.; Chia, L.; and Tsang, I. W. 2014. Region-based saliency detection and its application in object recognition. *IEEE TCSV* 24(5):769–779.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. S.; Berg, A. C.; and Li, F. 2015. Imagenet large scale visual recognition challenge. *IJCV* 115(3):211–252.

Shi, J.; Yan, Q.; Xu, L.; and Jia, J. 2016. Hierarchical image saliency detection on extended CSSD. *IEEE TPAMI* 38(4):717–729.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.

Tu, W.; He, S.; Yang, Q.; and Chien, S. 2016. Real-time salient object detection with a minimum spanning tree. In *CVPR*, pages 2334–2342.

Wang, J.; Jiang, H.; Yuan, Z.; Cheng, M.; Hu, X.; and Zheng, N. 2017. Salient object detection: A discriminative regional feature integration approach. *IJCV* 123(2):251–268.

Wang, L.; Wang, L.; Lu, H.; Zhang, P.; and Ruan, X. 2018. Salient object detection with recurrent fully convolutional networks. *IEEE TPAMI*.

Wei, Y.; Wen, F.; Zhu, W.; and Sun, J. 2012. Geodesic saliency using background priors. In *ECCV*, pages 29–42.

Wei, Y.; Liang, X.; Chen, Y.; Shen, X.; Cheng, M.; Feng, J.; Zhao, Y.; and Yan, S. 2017. STC: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE TIP* 39(11):2314–2320.

Zhang, L.; Yang, C.; Lu, H.; Ruan, X.; and Yang, M. 2017a. Ranking saliency. *IEEE TPAMI* 39(9):1892–1904.

Zhang, P.; Wang, D.; Lu, H.; Wang, H.; and Yin, B. 2017b. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, pages 212–221.

Zhang, P.; Liu, W.; Lu, H.; and Shen, C. 2018. Salient object detection by lossless feature reflection. In *IJCAI*, pages 1149–1155.

Zhao, R.; Ouyang, W.; Li, H.; and Wang, X. 2015. Saliency detection by multi-context deep learning. In *CVPR*, pages 1265–1274.

Zhu, W.; Liang, S.; Wei, Y.; and Sun, J. 2014. Saliency optimization from robust background detection. In *CVPR*, pages 2814–2821.