

SalGAN: visual saliency prediction with adversarial networks

Junting Pan^a, Cristian Canton-Ferrer^b, Kevin McGuinness^c, Noel E. O'Connor^c, Jordi Torres^d, Elisa Sayrol^a, Xavier Giro-i-Nieto^{a,**}

^aUniversitat Politècnica de Catalunya, Barcelona 08034, Catalonia/Spain

^bFacebook AML, Seattle, WA, United States of America

^cInsight Center for Data Analytics, Dublin City University, Dublin 9, Ireland

^dBarcelona Supercomputing Center, Barcelona 08034, Catalonia/Spain

ABSTRACT

Recent approaches for saliency prediction are generally trained with a loss function based on a single saliency metric. This could lead to low performance when evaluating with other saliency metrics. In this paper, we propose a novel data-driven metric based saliency prediction method, named SalGAN (Saliency GAN), trained with adversarial loss function. SalGAN consists of two networks: one predicts saliency maps from raw pixels of an input image; the other one takes the output of the first one to discriminate whether a saliency map is a predicted one or ground truth. By trying to make the predicted saliency map indistinguishable with the ground truth, SalGAN is expected to generate saliency maps that resembles the ground truth. Our experiments show that the adversarial training allows our model to obtain state-of-the-art performances across various saliency metrics.

© 2018 Elsevier Ltd. All rights reserved.

1. Introduction

Visual saliency describes the spatial locations in an image that attract the human attention. It is understood as a result of a bottom-up process where a human observer explores the image for a few seconds with no particular task in mind. Therefore, saliency prediction is indispensable for various machine vision tasks such as object recognition (Walther et al., 2002).

Visual saliency data are traditionally collected by eye-trackers (Judd et al., 2009a), and more recently with mouse clicks (Jiang et al., 2015) or webcams (Krafka et al., 2016). The salient points of the image are aggregated and convolved with a Gaussian kernel to obtain a saliency map. As a result, a gray-scale image or a heat map is generated to represent the probability of each corresponding pixel in the image to capture the human attention.

A lot of research effort has been made in designing an optimal loss function for saliency prediction. State-of-the-art methods (Huang et al., 2015) adopt saliency based metrics while others (Pan et al., 2016; Cornia et al., 2016; Jetley et al., 2016; Sun et al., 2017) use distance in saliency map space. How to choose

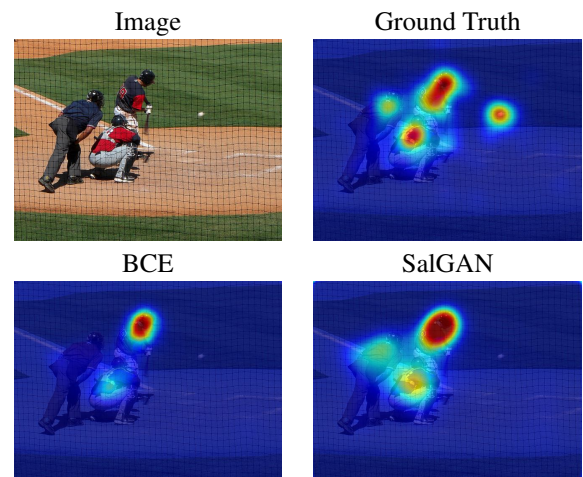


Fig. 1. Example of saliency map generation where the proposed system (SalGAN) outperforms a standard binary cross entropy (BCE) prediction model.

^{**}Corresponding author: Tel.: +34-934-015-769;
e-mail: xavier.giro@upc.edu (Xavier Giro-i-Nieto)

or design a best training loss is still an open problem. In addition, different saliency metrics diverge at defining the meaning of saliency maps, and there exist inconsistency with the model comparison. For instance, it has been pointed that the optimal metric for model optimization may depend on the final application (Bylinskii et al., 2016a).

To this end, instead of designing a tailored loss function, we introduce adversarial training for visual saliency prediction inspired by generative adversarial networks (GANs) (Goodfellow et al., 2014). We dub the proposed method as SalGAN. We focus on exploring the benefits of using such an adversarial loss to make the output saliency map not able to be distinguished from the real saliency maps. In GANs, training is driven by two competing agents: first, the *generator* synthesizing samples that match with the training data; second, the *discriminator* distinguishing between a real sample drawn directly from the training data and a fake one synthesized by the generator. In our case, this data distribution corresponds to pairs of real images and their corresponding visual saliency maps.

Specifically, SalGAN estimates the saliency map of an input image using a deep convolutional neural network (DCNN). As shown in the Figure 2 this network is initially trained with a binary cross entropy (BCE) loss over down-sampled versions of the saliency maps. The model is then refined with a discriminator network trained to solve a binary classification task between the saliency maps generated by SalGAN and the real ones used as ground truth. Our experiments show how adversarial training allows reaching state-of-the-art performance across different metrics when combined with a BCE content loss in a single-tower and single-task model.

To summarize, we investigate the introduction of the adversarial loss to the visual saliency learning. By introducing adversarial loss to the BCE saliency prediction model, we achieve the state-of-the-art performance in MIT300 and SALICON dataset for almost all the evaluation metrics

The remaining of the text is organized as follows. Section 2 reviews the state-of-the-art models for visual saliency prediction, discussing the loss functions they are based upon, their relations with the different metrics as well as their complexity in terms of architecture and training. Section 3 presents SalGAN, our deep convolutional neural network based on a convolutional encoder-decoder architecture, as well as the discriminator network used during its adversarial training. Section 4 describes the training process of SalGAN and the loss functions used. Section 5 includes the experiments and results of the presented techniques. Finally, Section 6 closes the paper by drawing the main conclusions.

Our results can be reproduced with the source code and trained models available at <https://imatge-upc.github.io/saliency-salgan-2017/>.

2. Related work

Saliency prediction has received interest by the research community for many years. Thus seminal works (Itti et al., 1998) proposed to predict saliency maps considering low-level features at multiple scales and combining them to form a

saliency map. (Harel et al., 2006), also starting from low-level feature maps, introduced a graph-based saliency model that defines Markov chains over various image maps, and treat the equilibrium distribution over map locations as activation and saliency values. (Judd et al., 2009b) presented a bottom-up, top-down model of saliency based not only on low but mid and high-level image features. (Borji, 2012) combined low-level features saliency maps of previous best bottom-up models with top-down cognitive visual features and learned a direct mapping from those features to eye fixations.

As in many other fields in computer vision, a number of deep learning solutions have very recently been proposed that significantly improve the performance. For example, the Ensemble of Deep Networks (eDN) (Vig et al., 2014) represented an early architecture that automatically learns the representations for saliency prediction, blending feature maps from different layers. In (Pan et al., 2016) two convolutional neural networks trained end-to-end for saliency prediction are compared, a lighter one designed and trained from scratch, and a second and deeper one pre-trained for image classification. DCNN have shown better results even when pre-trained with datasets build for other purposes. DeepGaze Kümmerer et al. (2015a) provided a deeper network using the well-know AlexNet (Krizhevsky et al., 2012), with pre-trained weights on Imagenet (Deng et al., 2009) and with a readout network on top whose inputs consisted of some layer outputs of AlexNet. The output of the network is blurred, center biased and converted to a probability distribution using a softmax. Huang et al. (Huang et al., 2015), in the so call SALICON net, obtained better results by using VGG rather than AlexNet or GoogleNet (Szegedy et al., 2015). In their proposal they considered two networks with fine and coarse inputs, whose feature maps outputs are concatenated.

Li et al. (Li and Yu, 2015) proposed a multi resolution convolutional neural network that is trained from image regions centered on fixation and non-fixation locations over multiple resolutions. Diverse top-down visual features can be learned in higher layers and bottom-up visual saliency can also be inferred by combining information over multiple resolutions. These ideas are further developed in they recent work called DSCLRCN Liu and Han (2018), where the proposed model learns saliency related local features on each image location in parallel and then learns to simultaneously incorporate global context and scene context to infer saliency. They incorporate a model to effectively learn long-term spatial interactions and scene contextual modulation to infer image saliency. Deep Gaze II (Kümmerer et al., 2017) sets the state of the art in the MIT300 dataset by combining features trained for image recognition with four layer of 1x1 convolutions. Both DSCLRCN and Deep Gaze II obtain excellent results in the benchmarks when combined with a center bias, which is not considered in SalGAN as the results are purely the results at inference time. MLNET (Cornia et al., 2016) proposes an architecture that combines features extracted at different levels of a DCNN. They introduce a loss function inspired by three objectives: to measure similarity with the ground truth, to keep invariance of predictive maps to their maximum and to give importance to

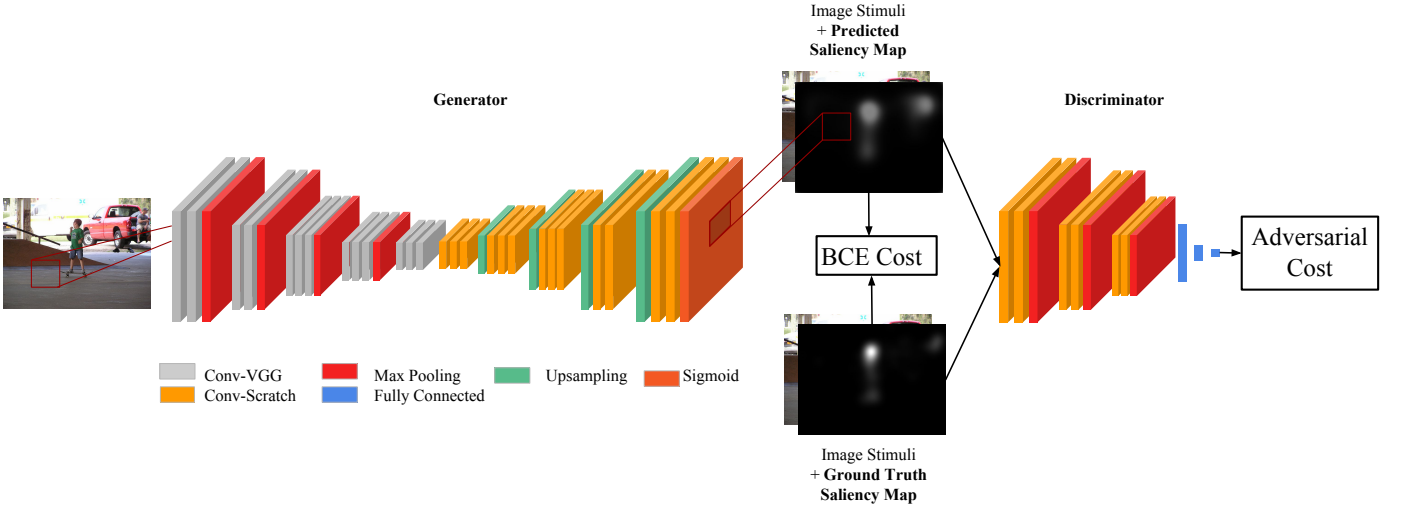


Fig. 2. Overall architecture of the proposed saliency system. The input for the saliency prediction network predicts an output saliency map given a natural image as input. Then, the pair of saliency and image is feed into the discriminator network. The output of the discriminator is a score that tells about whether the input saliency map is real or fake.

pixels with high ground truth fixation probability. In fact choosing an appropriate loss function has become an issue that can lead to improved results. Thus, another interesting contribution of (Huang et al., 2015) lies on minimizing loss functions based on metrics that are differentiable, such as NSS, CC, SIM and KL divergence to train the network (see Riche et al. (2013) and Kümmerer et al. (2015b) for the definition of these metrics. A thorough comparison of metrics can be found in (Bylinskii et al., 2016a)). In (Huang et al., 2015) KL divergence gave the best results. (Jetley et al., 2016) also tested loss functions based on probability distances, such as X2 divergence, total variation distance, KL divergence and Bhattacharyya distance by considering saliency map models as generalized Bernoulli distributions. The Bhattacharyya distance was found to give the best results.

In our work we present a network architecture that takes a different approach. By incorporating the high-level adversarial loss into the conventional saliency prediction training approach, the proposed method achieves the state-of-the-art performance in both MIT300 and SALICON datasets by a clear margin.

3. Architecture

The training of SalGAN is the result of two competing convolutional neural networks: a generator of saliency maps, which is SalGAN itself, and a discriminator network, which aims at distinguishing between the real saliency maps and those generated by SalGAN. This section provides details on the structure of both modules, the considered loss functions, and the initialization before beginning adversarial training. Figure 2 shows the architecture of the system.

3.1. Generator

The generator network, SalGAN, adopts a convolutional encoder-decoder architecture, where the encoder part includes

max pooling layers that decrease the size of the feature maps, while the decoder part uses upsampling layers followed by convolutional filters to construct an output that is the same resolution as the input.

The encoder part of the network is identical in architecture to VGG-16 (Simonyan and Zisserman, 2015), omitting the final pooling and fully connected layers. The network is initialized with the weights of a VGG-16 model trained on the ImageNet data set for object classification (Deng et al., 2009). Only the last two groups of convolutional layers in VGG-16 are modified during the training for saliency prediction, while the earlier layers remain fixed from the original VGG-16 model. We fix weights to save computational resources during training, even at the possible expense of some loss in performance.

The decoder architecture is structured in the same way as the encoder, but with the ordering of layers reversed, and with pooling layers being replaced by upsampling layers. Again, ReLU non-linearities are used in all convolution layers, and a final 1×1 convolution layer with sigmoid non-linearity is added to produce the saliency map. The weights for the decoder are randomly initialized. The final output of the network is a saliency map in the same size to input image.

The implementation details of SalGAN are presented in Table 1.

3.2. Discriminator

Table 2 gives the architecture and layer configuration for the discriminator. In short, the network is composed of six 3×3 kernel convolutions interspersed with three pooling layers ($\downarrow 2$), and followed by three fully connected layers. The convolution layers all use ReLU activations while the fully connected layers employ tanh activations, with the exception of the final layer, which uses a sigmoid activation.

layer	depth	kernel	stride	pad	activation
conv1_1	64	1×1	1	1	ReLU
conv1_2	64	3×3	1	1	ReLU
pool1		2×2	2	0	-
conv2_1	128	3×3	1	1	ReLU
conv2_2	128	3×3	1	1	ReLU
pool2		2×2	2	0	-
conv3_1	256	3×3	1	1	ReLU
conv3_2	256	3×3	1	1	ReLU
conv3_3	256	3×3	1	1	ReLU
pool3		2×2	2	0	-
conv4_1	512	3×3	1	1	ReLU
conv4_2	512	3×3	1	1	ReLU
conv4_3	512	3×3	1	1	ReLU
pool4		2×2	2	0	-
conv5_1	512	3×3	1	1	ReLU
conv5_2	512	3×3	1	1	ReLU
conv5_3	512	3×3	1	1	ReLU
conv6_1	512	3×3	1	1	ReLU
conv6_2	512	3×3	1	1	ReLU
conv6_3	512	3×3	1	1	ReLU
upsample6		2×2	2	0	-
conv7_1	512	3×3	1	1	ReLU
conv7_2	512	3×3	1	1	ReLU
conv7_3	512	3×3	1	1	ReLU
upsample7		2×2	2	0	-
conv8_1	256	3×3	1	1	ReLU
conv8_2	256	3×3	1	1	ReLU
conv8_3	256	3×3	1	1	ReLU
upsample8		2×2	2	0	-
conv9_1	128	3×3	1	1	ReLU
conv9_2	128	3×3	1	1	ReLU
upsample9		2×2	2	0	-
conv10_1	64	3×3	1	1	ReLU
conv10_2	64	3×3	1	1	ReLU
output	1	1×1	1	0	Sigmoid

Table 1. Architecture of the generator network.

layer	depth	kernel	stride	pad	activation
conv1_1	3	1×1	1	1	ReLU
conv1_2	32	3×3	1	1	ReLU
pool1		2×2	2	0	-
conv2_1	64	3×3	1	1	ReLU
conv2_2	64	3×3	1	1	ReLU
pool2		2×2	2	0	-
conv3_1	64	3×3	1	1	ReLU
conv3_2	64	3×3	1	1	ReLU
pool3		2×2	2	0	-
fc4	100	-	-	-	tanh
fc5	2	-	-	-	tanh
fc6	1	-	-	-	sigmoid

Table 2. Architecture of the discriminator network.

4. Training

The filter weights in SalGAN have been trained over a perceptual loss (Johnson et al., 2016) resulting from combining a content and adversarial loss. The content loss follows a classic approach in which the predicted saliency map is pixel-wise compared with the corresponding one from ground truth. The adversarial loss depends of the real/synthetic prediction of the discriminator over the generated saliency map.

4.1. Content loss

The content loss is computed in a per-pixel basis, where each value of the predicted saliency map is compared with its corresponding peer from the ground truth map. Given an image I of dimensions $N = W \times H$, we represent the saliency map S as vector of probabilities, where S_j is the probability of pixel I_j being fixated. A content loss function $\mathcal{L}(S, \hat{S})$ is defined between the predicted saliency map \hat{S} and its corresponding ground truth S .

The first considered content loss is mean squared error (MSE) or Euclidean loss, defined as:

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{j=1}^N (S_j - \hat{S}_j)^2. \quad (1)$$

In our work, MSE is used as a baseline reference, as it has been adopted directly or with some variations in other state of the art solutions for visual saliency prediction (Pan et al., 2016; Cornia et al., 2016).

Solutions based on MSE aim at maximizing the peak signal-to-noise ratio (PSNR). These works tend to filter high spatial frequencies in the output, favoring this way blurred contours. MSE corresponds to computing the Euclidean distance between the predicted saliency and the ground truth.

Ground truth saliency maps are normalized so that each value is in the range $[0, 1]$. Saliency values can therefore be interpreted as estimates of the probability that a particular pixel is attended by an observer. It is tempting to therefore induce a multinomial distribution on the predictions using a softmax on the final layer. Clearly, however, more than a single pixel may be attended, making it more appropriate to treat each predicted value as independent of the others. We therefore propose to apply an element-wise sigmoid to each output in the final layer so that the pixel-wise predictions can be thought of as probabilities for independent binary random variables. An appropriate loss in such a setting is the binary cross entropy, which is the average of the individual binary cross entropies (BCE) across all pixels:

$$\mathcal{L}_{BCE} = -\frac{1}{N} \sum_{j=1}^N (S_j \log(\hat{S}_j) + (1 - S_j) \log(1 - \hat{S}_j)). \quad (2)$$

4.2. Adversarial loss

Generative adversarial networks (GANs) (Goodfellow et al., 2014) are commonly used to generate images with realistic statistical properties. The idea is to simultaneously fit two parametric functions. The first of these functions, known as the generator, is trained to transform samples from a simple distribution (e.g. Gaussian) into samples from a more complicated

distribution (e.g. natural images). The second function, the discriminator, is trained to distinguish between samples from the true distribution and generated samples. Training proceeds alternating between training the discriminator using generated and real samples, and training the generator, by keeping the discriminator weights constant and backpropagating the error through the discriminator to update the generator weights.

The saliency prediction problem has some important differences from the above scenario. First, the objective is to fit a deterministic function that predict realistic saliency values from images, rather than realistic images from random noise. As such, in our case the input to the generator (saliency prediction network) is not random noise but an image. Second, the input image that a saliency map corresponds to is essential, due the fact the goal is not only to have the two saliency maps becoming indistinguishable but with the condition that they both correspond the same input image. We therefore include both the image and saliency map as inputs to the discriminator network. Finally, when using generative adversarial networks to generate realistic images, there is generally no ground truth to compare against. In our case, however, the corresponding ground truth saliency map is available. When updating the parameters of the generator function, we found that using a loss function that is a combination of the error from the discriminator and the cross entropy with respect to the ground truth improved the stability and convergence rate of the adversarial training. The final loss function for the saliency prediction network during adversarial training can be formulated as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{BCE} + L(D(I, \hat{S}), 1). \quad (3)$$

where L is the binary cross entropy loss, and 1 is the target category of real samples and 0 for the category of fake (predicted) sample. Here, instead of minimizing $-L(D(I, S), 0)$, we optimize $L(D(I, S), 1)$ which provides stronger gradient, similar to (Goodfellow et al., 2014). $D(I, \hat{S})$ is the probability of fooling the discriminator network, so that the loss associated to the saliency prediction network will grow more when chances of fooling the discriminator are lower. During the training of the discriminator, no content loss is available and the loss function is:

$$\mathcal{L}_D = L(D(I, S), 1) + L(D(I, \hat{S}), 0). \quad (4)$$

At train time, we first bootstrap the saliency prediction network function by training for 15 epochs using only BCE, which is computed with respect to the down-sampled output and ground truth saliency. After this, we add the discriminator and begin adversarial training. The input to the discriminator network is an RGBS image of size $256 \times 192 \times 4$ containing both the source image channels and (predicted or ground truth) saliency.

We train the networks on the 15,000 images from the SALICON training set using a batch size of 32. During the adversarial training, we alternate the training of the saliency prediction network and discriminator network after each iteration (batch). We used L2 weight regularization (i.e. weight decay) when training both the generator and discriminator ($\lambda = 1 \times 10^{-4}$). We used AdaGrad for optimization, with an initial learning rate of 3×10^{-4} .

	sAUC \uparrow	AUC-B \uparrow	NSS \uparrow	CC \uparrow	IG \uparrow
BCE	0.752	0.825	2.473	0.761	0.712
BCE/2	0.750	0.820	2.527	0.764	0.592
BCE/4	0.755	0.831	2.511	0.763	0.825
BCE/8	0.754	0.827	2.503	0.762	0.831

Table 3. Impact of downsampled saliency maps at (15 epochs) evaluated over SALICON validation. BCE/ x refers to a downsample factor of $1/x$ over a saliency map of 256×192 .

5. Experiments

The presented SalGAN model for visual saliency prediction was assessed and compared from different perspectives. First, the impact of using BCE and the downsampled saliency maps are assessed. Second, the gain of the adversarial loss is measured and discussed, both from a quantitative and a qualitative point of view. Finally, the performance of SalGAN is compared to published works to compare its performance with the current state-of-the-art. The experiments aimed at finding the best configuration for SalGAN were run using the *train* and *validation* partitions of the SALICON dataset (Jiang et al., 2015). This is a large dataset built by collecting mouse clicks on a total of 20,000 images from the Microsoft Common Objects in Context (MS-CoCo) dataset (Lin et al., 2014). We have adopted this dataset for our experiments because it is the largest one available for visual saliency prediction. In addition to SALICON, we also present results on MIT300, the benchmark with the largest amount of submissions.

5.1. Non-adversarial training

The two content losses presented in content loss section, MSE and BCE, were compared to define a baseline upon which we later assess the impact of the adversarial training. The two first rows of Table 4 shows how a simple change from MSE to BCE brings a consistent improvement in all metrics. This improvement suggests that treating saliency prediction as multiple binary classification problem is more appropriate than treating it as a standard regression problem, in spite of the fact that the target values are not binary. Minimizing cross entropy is equivalent to minimizing the KL divergence between the predicted and target distributions, which is a reasonable objective if both predictions and targets are interpreted as probabilities.

Based on the superior BCE-based loss compared with MSE, we also explored the impact of computing the content loss over downsampled versions of the saliency map. This technique reduces the required computational resources at both training and test times and, as shown in Table 3, not only does it not decrease performance, but it can actually improve it. Given this results, we chose to train SalGAN on saliency maps downsampled by a factor $1/4$, which in our architecture corresponds to saliency maps of 64×48 .

5.2. Adversarial gain

The adversarial loss was introduced after estimating the value of the hyperparameter α in Equation 3 by maximizing the most general metric, Information Gain (IG). As shown in Figure 3,

the search was performed on logarithmic scale, and we achieved the best performance for $\alpha = 0.005$.

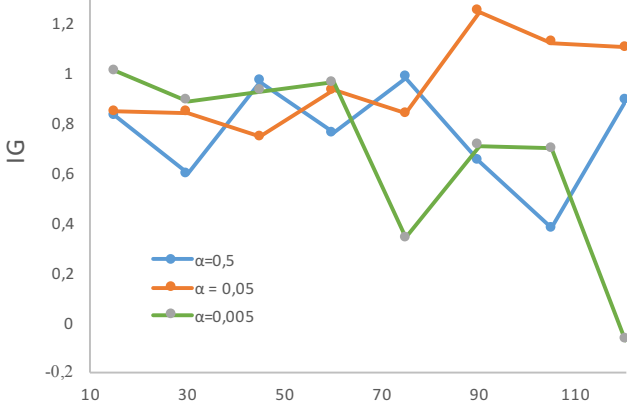


Fig. 3. SALICON validation set Information Gain for different hyper parameter α on varying numbers of epochs.

The Information Gains (IG) of SalGAN for different values of the hyper parameter α are compared in Figure 3. The search for finding an optimal hyper parameter α is performed on logarithmic scale, and we achieved the best performance for $\alpha = 0.005$.

The gain achieved by introducing the adversarial loss into the perceptual loss was assessed by using BCE as a content loss and feature maps of 68×48 . The first row of results in Table 4 refers to a baseline defined by training SalGAN with the BCE content loss for 15 epochs only. Later, two options are considered: 1) training based on BCE only (2nd row), or 2) introducing the adversarial loss (3rd and 4th row).

Figure 4 compares validation set accuracy metrics for training with combined GAN and BCE loss versus a BCE alone as the number of epochs increases. In the case of the AUC metrics (Judd and Borji), increasing the number of epochs does not lead to significant improvements when using BCE alone. The combined BCE/GAN loss however, continues to improve performance with further training. After 100 and 120 epochs, the combined GAN/BCE loss shows substantial improvements over BCE for five of six metrics.

The single metric for which Adversarial training fails to improve performance is normalized scanpath saliency (NSS). The reason for this may be that GAN training tends to produce a smoother and more spread out estimate of saliency, which better matches the statistical properties of real saliency maps, but may increase the false positive rate. As noted in (Bylinskii et al., 2016a), NSS is very sensitive to such false positives. The impact of increased false positives depends on the final application. In applications where the saliency map is used as a multiplicative attention model (e.g. in retrieval applications, where spatial features are importance weighted), false positives are often less important than false negatives, since while the former includes more distractors, the latter removes potentially useful features. Note also that NSS is differentiable, so could potentially be optimized directly when important for a particular ap-

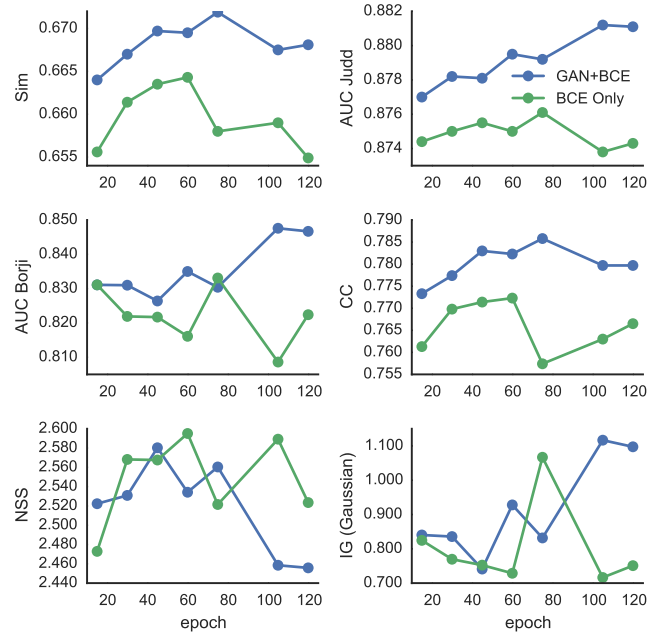


Fig. 4. SALICON validation set accuracy metrics for Adversarial+BCE vs BCE on varying numbers of epochs. AUC shuffled is omitted as the trend is identical to that of AUC Judd.

	sAUC \uparrow	AUC-B \uparrow	NSS \uparrow	CC \uparrow	IG
MSE	0.728	0.820	1.680	0.708	0.628
BCE	0.753	0.825	2.562	0.772	0.824
BCE/4	0.757	0.833	2.580	0.772	1.067
GAN/4	0.773	0.859	2.560	0.786	1.243

Table 4. Best results through epochs obtained with non-adversarial (MSE and BCE) and adversarial training. BCE/4 and GAN/4 refer to downsampled saliency maps. Saliency maps assessed on SALICON validation.

plication.

5.3. Comparison with the state-of-the-art

SalGAN is compared in Table 5 to several other algorithms from the state-of-the-art. The comparison is based on the evaluations run by the organizers of the SALICON and MIT300 benchmarks on a test dataset whose ground truth is not public. The two benchmarks offer complementary features: while SALICON is a much larger dataset with 5,000 test images, MIT300 has attracted the participation of many more researchers. In both cases, SalGAN was trained using 15,000 images contained in the training (10,000) and validation (5,000) partitions of the SALICON dataset. Notice that while both datasets aim at capturing visual saliency, the acquisition of data differed, as SALICON ground truth was generated based on crowdsourced mouse clicks, while the MIT300 was built with eye trackers on a limited and controlled group of users. Table 5 compares SalGAN with other contemporary works that have used SALICON and MIT300 datasets. SalGAN presents very competitive results in both datasets, as it improves or equals the performance of all other models in at least one metric.

SALICON (test)	AUC-J \uparrow	Sim \uparrow	EMD \downarrow	AUC-B \uparrow	sAUC \uparrow	CC \uparrow	NSS \uparrow	KL \downarrow
DSCLRCN (Liu and Han, 2018)	-	-	-	0.884	0.776	0.831	3.157	-
SalGAN	-	-	-	0.884	0.772	0.781	2.459	-
ML-NET (Cornia et al., 2016)	-	-	-	(0.866)	(0.768)	(0.743)	2.789	-
SalNet (Pan et al., 2016)	-	-	-	(0.858)	(0.724)	(0.609)	(1.859)	-
MIT300	AUC-J \uparrow	Sim \uparrow	EMD \downarrow	AUC-B \uparrow	sAUC \uparrow	CC \uparrow	NSS \uparrow	KL \downarrow
Humans	0.92	1.00	0.00	0.88	0.81	1.0	3.29	0.00
Deep Gaze II Kümmerer et al. (2017)	(0.84)	(0.43)	(4.52)	(0.83)	0.77	(0.45)	(1.16)	(1.04)
DSCLRCN (Liu and Han, 2018)	0.87	0.68	2.17	(0.79)	0.72	0.80	2.35	0.95
SALICON (Huang et al., 2015)	0.87	(0.60)	(2.62)	0.85	0.74	0.74	2.12	0.54
SalGAN	0.86	0.63	2.29	0.81	0.72	0.73	2.04	1.07
PDP (Jetley et al., 2016)	(0.85)	(0.60)	(2.58)	(0.80)	0.73	(0.70)	2.05	0.92
ML-NET (Cornia et al., 2016)	(0.85)	(0.59)	(2.63)	(0.75)	(0.70)	(0.67)	2.05	(1.10)
Deep Gaze I (Kümmerer et al., 2015a)	(0.84)	(0.39)	(4.97)	0.83	(0.66)	(0.48)	(1.22)	(1.23)
SalNet (Pan et al., 2016)	(0.83)	(0.52)	(3.31)	0.82	(0.69)	(0.58)	(1.51)	0.81
BMS (Zhang and Sclaroff, 2013)	(0.83)	(0.51)	(3.35)	0.82	(0.65)	(0.55)	(1.41)	0.81

Table 5. Comparison of SalGAN with other state-of-the-art solutions on the SALICON (test) and MIT300 benchmarks. Values in brackets correspond to performances worse than SalGAN.

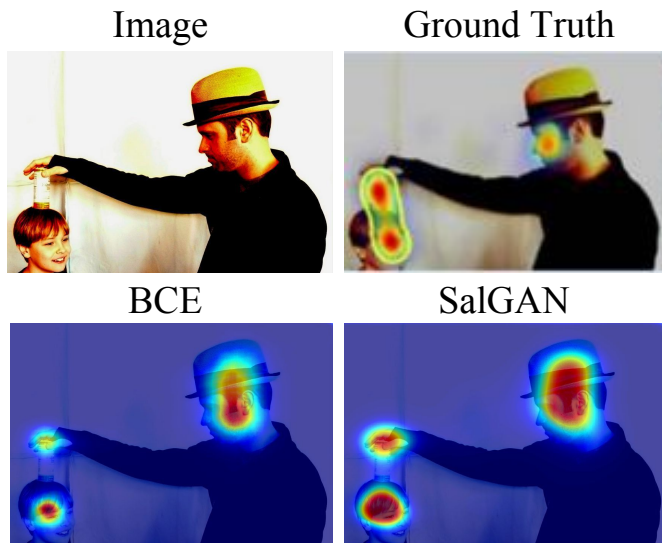


Fig. 5. Example images from MIT300 containing salient region (marked in yellow) that is often missed by computational models, and saliency map estimated by SalGAN.

5.4. Qualitative results

The impact of adversarial training has also been explored from a qualitative perspective by observing the resulting saliency maps. Figure 5 shows one example from the MIT300 dataset, highlighted in (Bylinskii et al., 2016b) as being particular challenges for existing saliency algorithms. The areas highlighted in yellow in the images on the left are regions that are typically missed by algorithms. In this example, we see that SalGAN successfully detects the often missed hand of the magician and face of the boy as being salient.

Figure 6 illustrates the effect of adversarial training on the statistical properties of the generated saliency maps. Shown are two close up sections of a saliency map from cross entropy training (left) and adversarial training (right). Training on BCE alone produces saliency maps that while they may be locally

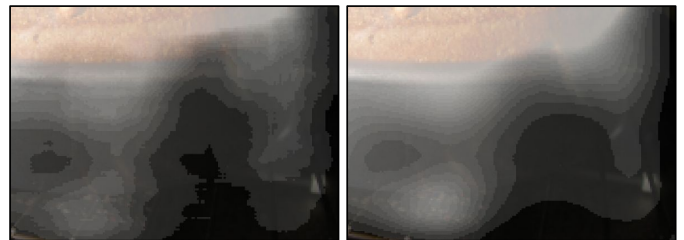


Fig. 6. Close-up comparison of output from training on BCE loss vs combined BCE/Adversarial loss. Left: saliency map from network trained with BCE loss. Right: saliency map from proposed adversarial training.

consistent with the ground truth, are often less smooth and have complex level sets. Adversarial training on the other hand produces much smoother and simpler level sets. Finally, Figure 7 shows some qualitative results comparing the results from training with BCE and BCE/Adversarial against the ground truth for images from the SALICON validation set.

6. Conclusions

To the best of our knowledge, this is the first work that proposes an adversarial-based approach to saliency prediction and has shown how adversarial training over a deep convolutional neural network can achieve state-of-the-art performance with a simple encoder-decoder architecture. A BCE-based content loss was shown to be effective for both initializing the saliency prediction network, and as a regularization term for stabilizing adversarial training. Our experiments showed that adversarial training improved all bar one saliency metric when compared to further training on cross entropy alone.

It is worth pointing out that although we use a VGG-16 based encoder-decoder model as the saliency prediction network in this paper, the proposed adversarial training approach is generic and could be applied to improve the performance of other saliency models.

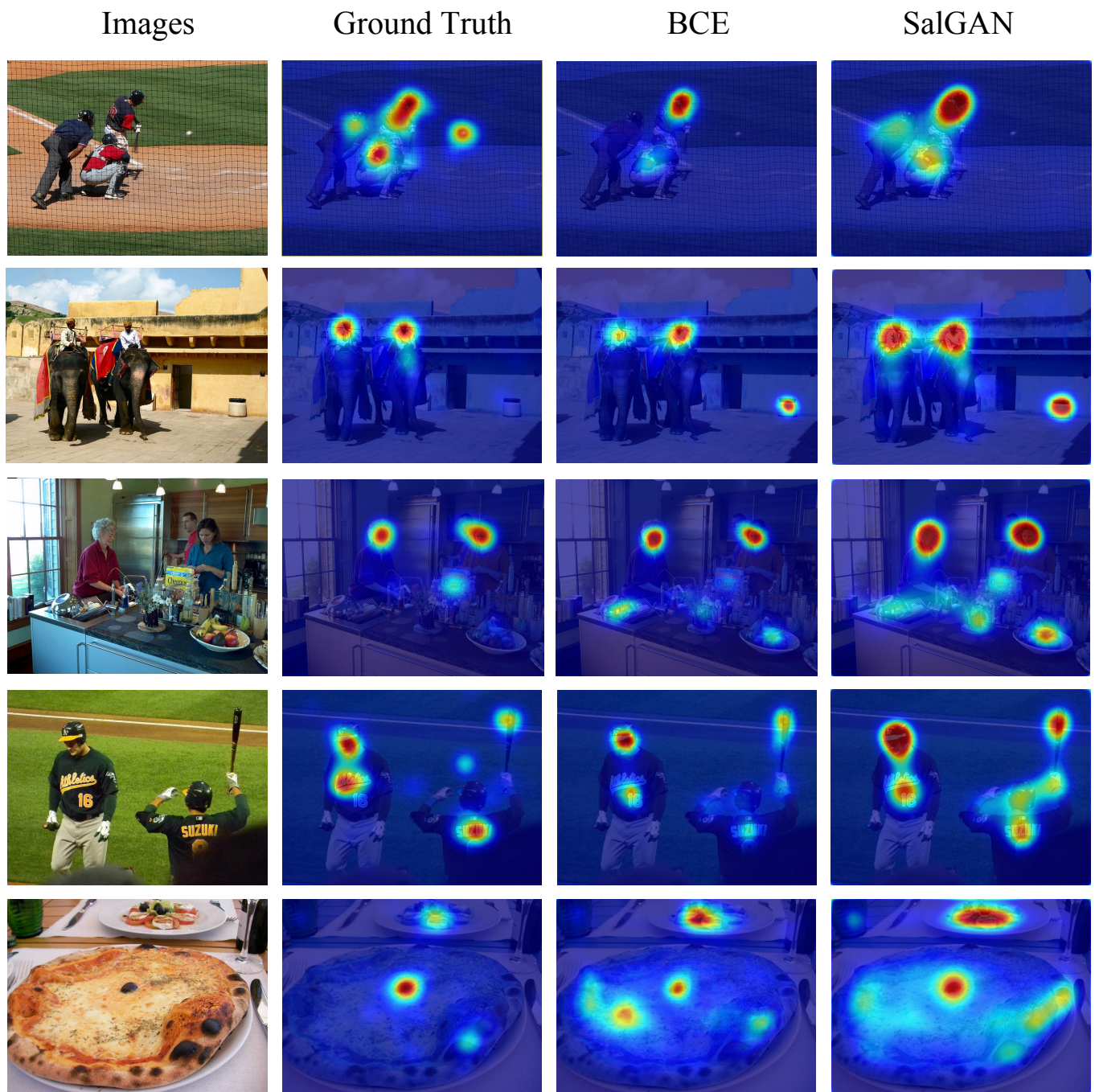


Fig. 7. Qualitative results of SalGAN on the SALICON validation set. SalGAN predicts well those high salient regions which are missed by BCE model. Saliency maps of BCE model are very localized in a few salient regions, they tend to fail when the number of salient regions increases.

Acknowledgments

The Image Processing Group at UPC is supported by the project TEC2016-75976-R, funded by the Spanish Ministerio de Economía y Competitividad and the European Regional Development Fund (ERDF). This material is based upon works supported by Science Foundation Ireland under Grant No 15/SIRG/3283. We gratefully acknowledge the support of NVIDIA Corporation for the donation of GPUs used in this work.

References

- Borji, A., 2012. Boosting bottom-up and top-down visual features for saliency estimation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F., 2016a. What do different evaluation metrics tell us about saliency models? ArXiv preprint:1610.01563.
- Bylinskii, Z., Recasens, A., Borji, A., Oliva, A., Torralba, A., Durand, F., 2016b. Where should saliency models look next?, in: European Conference on Computer Vision (ECCV).
- Cornia, M., Baraldi, L., Serra, G., Cucchiara, R., 2016. A deep multi-level network for saliency prediction, in: International Conference on Pattern Recognition (ICPR).
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in Neural Information Processing Systems, pp. 2672–2680.
- Harel, J., Koch, C., Perona, P., 2006. Graph-based visual saliency, in: Neural Information Processing Systems (NIPS).
- Huang, X., Shen, C., Boix, X., Zhao, Q., 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks, in: IEEE International Conference on Computer Vision (ICCV).
- Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 1254–1259.
- Jetley, S., Murray, N., Vig, E., 2016. End-to-end saliency mapping via probability distribution prediction, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Jiang, M., Huang, S., Duan, J., Zhao, Q., 2015. Salicon: Saliency in context, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision (ECCV).
- Judd, T., Ehinger, K., Durand, F., Torralba, A., 2009a. Learning to predict where humans look, in: Computer Vision, 2009 IEEE 12th international conference on, IEEE, pp. 2106–2113.
- Judd, T., Ehinger, K., Durand, F., Torralba, A., 2009b. Learning to predict where humans look, in: IEEE International Conference on Computer Vision (ICCV).
- Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A., 2016. Eye tracking for everyone, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.
- Kümmerer, M., Theis, L., Bethge, M., 2015a. DeepGaze I: Boosting saliency prediction with feature maps trained on ImageNet, in: International Conference on Learning Representations (ICLR).
- Kümmerer, M., Theis, L., Bethge, M., 2015b. Information-theoretic model comparison unifies saliency metrics. Proceedings of the National Academy of Sciences (PNAS) 112, 16054–16059.
- Kümmerer, M., Wallis, T.S., Gatys, L.A., Bethge, M., 2017. Understanding low-and high-level contributions to fixation prediction, in: 2017 IEEE International Conference on Computer Vision, pp. 4799–4808.
- Li, G., Yu, Y., 2015. Visual saliency based on multiscale deep features, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, L., 2014. Microsoft COCO: Common objects in context, in: European Conference on Computer Vision (ECCV).
- Liu, N., Han, J., 2018. A deep spatial contextual long-term recurrent convolutional network for saliency detection. IEEE Transactions on Image Processing 27, 3264–3274.
- Pan, J., Sayrol, E., Giró-i Nieto, X., McGuinness, K., O’Connor, N.E., 2016. Shallow and deep convolutional networks for saliency prediction, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Riche, N., M., D., Mancas, M., Gosselin, B., Dutoit, T., 2013. Saliency and human fixations. state-of-the-art and study comparison metrics, in: IEEE International Conference on Computer Vision (ICCV).
- Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: International Conference on Learning Representations (ICLR).
- Sun, X., Huang, Z., Yin, H., Shen, H.T., 2017. An integrated model for effective saliency prediction., in: AAAI, pp. 274–281.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P.R.S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Vig, E., Dorr, M., Cox, D., 2014. Large-scale optimization of hierarchical features for saliency prediction in natural images, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
- Walther, D., Itti, L., Riesenhuber, M., Poggio, T., Koch, C., 2002. Attentional selection for object recognition-a gentle way. Lecture Notes in Computer Science 2525, 472–479.
- Zhang, J., Sclaroff, S., 2013. Saliency detection: a boolean map approach, in: IEEE International Conference on Computer Vision (ICCV).