

---

# Big GANs Are Watching You: Towards Unsupervised Object Segmentation with Off-the-Shelf Generative Models

---

**Andrey Voynov**

Yandex

an.voynov@yandex.ru

**Stanislav Morozov**

Yandex

Lomonosov Moscow State University

stanis-morozov@yandex.ru

**Artem Babenko**

Yandex

National Research University

Higher School of Economics

artem.babenko@phystech.edu

## Abstract

Since collecting pixel-level groundtruth data is expensive, unsupervised visual understanding problems are currently an active research topic. In particular, several recent methods based on generative models have achieved promising results for object segmentation and saliency detection. However, since generative models are known to be unstable and sensitive to hyperparameters, the training of these methods can be challenging and time-consuming.

In this work, we introduce an alternative, much simpler way to exploit generative models for unsupervised object segmentation. First, we explore the latent space of the BigBiGAN — the state-of-the-art unsupervised GAN, which parameters are publicly available. We demonstrate that object saliency masks for GAN-produced images can be obtained automatically with BigBiGAN. These masks then are used to train a discriminative segmentation model. Being very simple and easy-to-reproduce, our approach provides competitive performance on common benchmarks in the unsupervised scenario. Code is available online<sup>1</sup>.

## 1 Introduction

Deep convolutional models are a core instrument for visual understanding problems, including object localization[1, 2], saliency detection [3], segmentation[4] and others. Deep models, however, require a large amount of high-quality training data to fit a huge number of learnable parameters. In practice, obtaining groundtruth pixel-level labeling is expensive, since it requires labor-intensive human efforts. Therefore, much research attention has currently focused on weakly-supervised and unsupervised approaches for challenging pixel-level tasks, such as segmentation[5, 6, 7, 8].

An emerging line of research on unsupervised segmentation exploits generative models as a tool for image decomposition. Namely, recent works [7, 8] have designed training protocols that include generative adversarial networks (GANs), to solve the foreground object segmentation without human labels. Given the promising results and the fact that the GANs' performance is steadily improving, this research direction will likely develop in the future.

---

<sup>1</sup><https://github.com/anvoynov/BigGANsAreWatching>

In practice, however, training high-quality generative models is challenging. This is especially the case for GANs, which training can be both time-consuming and unstable. Moreover, the models in [7, 8] typically include a large number of hyperparameters that can be tricky to tune, especially in the completely unsupervised scenario when labeled validation set is not available.

To this end, we propose an alternative way to exploit GANs for unsupervised segmentation, which does not train a separate generative model for each task. Instead, we use a publicly available pretrained GAN to generate synthetic images equipped with segmentation masks, which can be obtained automatically. In more detail, we explore the latent space of the publicly available BigBiGAN model [9], which is an unsupervised GAN trained on the Imagenet[10]. With the recent unsupervised technique [11], we demonstrate that manipulations in the BigBiGAN latent space allow to distinguish object/background pixels in the generated images, providing decent segmentation masks. These masks are then used to supervise a discriminative U-Net model [12], which is much easier to train. As another advantage, our approach also provides a straightforward way to tune hyperparameters. Since an amount of synthetic data is unlimited, its hold-out subset can be used as validation.

Our work confirms the promise of using GANs to produce synthetic training data, which is a long-standing goal of research on generative modeling. In extensive experiments, we show that the approach often outperforms the existing unsupervised alternatives for object segmentation and saliency detection. Furthermore, our approach performs on par with weakly-supervised methods for object localization, despite being completely unsupervised.

The main contributions of our paper are the following:

1. We introduce an alternative line of research on using GANs for unsupervised object segmentation. In a nutshell, we advocate the usage of high-quality synthetic data produced by BigBiGAN, which can provide high-quality saliency masks for generated images.
2. We compare our method to existing approaches and achieve a new state-of-the-art in most operating points. Given its simplicity, the method can serve as a useful baseline in the future.
3. We demonstrate a novel unsupervised scenario, where GAN-produced imagery becomes a useful source of training data for supervised computer vision models.

## 2 Related work

In this paper, we address the binary object segmentation problem, i.e, for each image pixel we aim to predict if it belongs to the object or to the background. In the literature this setup is typically referred to as saliency detection[3] and foreground object segmentation[7, 8]. While most prior works operate in fully-supervised or weakly-supervised regimes, we focus on the most challenging unsupervised scenario, for which only a few approaches have been developed.

**Existing unsupervised approaches.** Before a rise of deep learning models, a large number of “shallow” unsupervised techniques were developed [13, 14, 15, 16, 17, 18]. These earlier techniques were mostly based on hand-crafted features and heuristics, e.g., color contrast[17] or certain background priors [18]. Often these approaches also utilize traditional computer vision routines, such as superpixels[19, 20], object proposals[21], CRF[22]. These heuristics, however, are not completely learned from data, and the corresponding methods are inferior to the more recent “deep” approaches.

Regarding unsupervised deep models, several works have recently been proposed by the saliency detection community[23, 24, 25, 26]. Their main idea is to combine or fuse the predictions of several heuristic saliency methods, typically using them as a source of noisy groundtruth for deep CNN models. However, these methods are not completely unsupervised, since they typically rely on the pretrained classification or segmentation networks. In contrast, in this work, we focus on the methods that do not require any source of external supervision.

**Generative models for object segmentation.** The recent line of completely unsupervised methods [7, 8] employs generative modeling to decompose the image into the object and the background. In a nutshell, these methods exploit the idea that the object location or appearance can be perturbed without affecting the image realism. This inductive bias is formalized in the training protocols [7, 8], which include learning of GANs. Therefore, for each new segmentation task, one has to perform adversarial learning, which is known to be unstable, time-consuming, and sensitive to hyperparameters.

In contrast, our approach avoids these disadvantages, being much simpler and easier to reproduce. In essence, we propose to use the “inner knowledge” of the off-the-shelf large-scale GAN to produce the saliency masks for synthetic images and use them as a supervision for discriminative models.

**Latent spaces of large-scale GANs.** Our study is partially inspired by the recent findings from [11]. This work introduces an unsupervised technique that discovers the directions in the GAN latent space corresponding to interpretable image transformations. Among its findings, [11] demonstrates that the large-scale conditional GAN (BigGAN [27]) possesses a “background removal” direction that can be used to obtain saliency masks. However, this direction was discovered only for BigGAN that was trained under the supervision from the image class labels. For unconditional GANs, such a direction was not discovered in [11], hence, it is not clear if the supervision from the class labels is necessary for the GAN latent space “to understand” what pixels belong to object/background. In this paper, we demonstrate that this supervision is not necessary, therefore, even completely unsupervised GANs can serve as an excellent source of synthetic data for object segmentation.

### 3 Method

#### 3.1 Exploring the BigBiGAN latent space.

The main component of our method is the recent BigBiGAN model [9]. BigBiGAN is the state-of-the-art generative adversarial network trained on the Imagenet [10] without labels and its parameters are available online<sup>2</sup>. The BigBiGAN generator  $G$  maps the samples  $z \sim \mathcal{N}(0, \mathbb{I})$  from the latent space  $\mathbb{R}^{120}$  into the image space  $G : z \rightarrow I$ . BigBiGAN is also equipped with an encoder  $E : I \rightarrow z$  that was trained jointly with the generator and maps images to the latent space. In this section, we explore the BigBiGAN latent space to investigate if its properties can be useful for downstream tasks.

A very recent paper [11] has introduced an unsupervised technique that identifies interpretable directions in the latent space of a pretrained GAN. By moving a latent code  $z$  in these directions, one can achieve different image transformations, such as image zooming or translation. Formally, given an image corresponding to a latent code  $z$ , one can modify it via shifting the code in an interpretable direction  $h$ . Then a modified image  $G(z+h)$  can be generated. Importantly,  $h$  operates consistently over the whole latent space, i.e. for all  $z$ , shifting results in the same type of transformation. As the first step of our study, we apply the technique from [11] to the BigBiGAN generator to explore the potential of its latent space. In a nutshell, [11] seeks to learn  $K$  directions in the latent space  $h_1, \dots, h_K$  such that the effects of the corresponding image transformations are “disentangled”. More formally, the sets of pairs  $\{G(z), G(z+h_i) | z \sim \mathcal{N}(0, \mathbb{I})\}$  for different  $i=1, \dots, K$  are easy to distinguish from each other by a CNN classifier, which is trained jointly with  $h_1, \dots, h_K$ .

We use the authors’ implementation<sup>3</sup> with default hyperparameters and the number of directions  $K=120$ . After learning converged, we inspect the directions manually and filter out only the directions that are interpretable. Several directions revealed by the procedure are provided in Figure 1. Compared to the results from [11] for the “supervised” conditional BigGAN, the BigBiGAN latent space does not possess any directions that have clear “background removal” effect. However, one of the directions has an effect that can be used to distinguish between object and background pixels. The corresponding transformation “Saliency lighting” is presented on Figure 1 and we refer to this direction as  $h_{bg}$ . As one can see, moving in this direction makes the object pixels lighter, while the background pixels become darker. Therefore, despite BigBiGAN is completely unsupervised, its latent space can be used to obtain saliency masks for generated images. Technically, we produce a binary saliency mask  $M$  for an image  $G(z)$  by comparing its intensity with the “shifted” image  $M=[G(z+h_{bg}) > G(z)]$  after grayscale conversion. As a shift magnitude, we always use  $\|h_{bg}\|=5$ .

#### 3.2 Improving saliency masks.

Here we describe a few tricks increasing the quality of the masks for the particular segmentation task.

**Adaptation to the particular segmentation task.** In the scheme above the latent codes are sampled from the standard Gaussian distribution  $z \sim \mathcal{N}(0, \mathbb{I})$ . To make the distribution of generated images closer to the particular dataset at hand  $\mathcal{I}=\{I_1, \dots, I_N\}$ , we aim to sample  $z$  from the latent space regions that are close to the latent codes of  $\mathcal{I}$ . To this end, we use the BigBiGAN encoder to compute

<sup>2</sup> <https://tfhub.dev/deepmind/bigbigan-resnet50/1>

<sup>3</sup> <https://github.com/anvoynov/GANLatentDiscovery>

Figure 1: Examples of interpretable directions discovered in the BigBiGAN latent space.

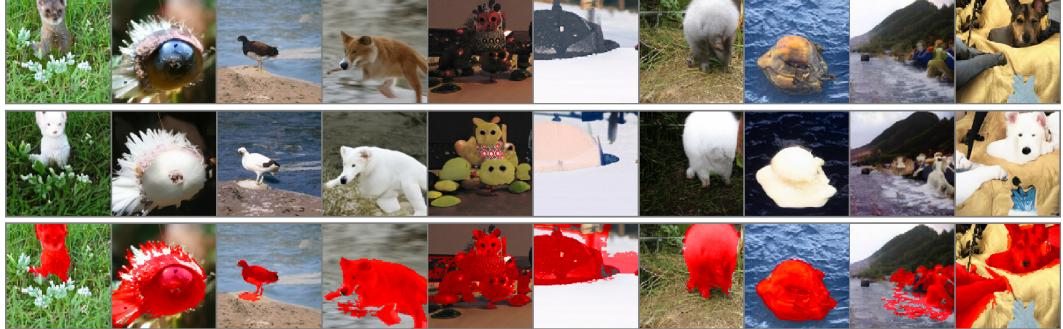
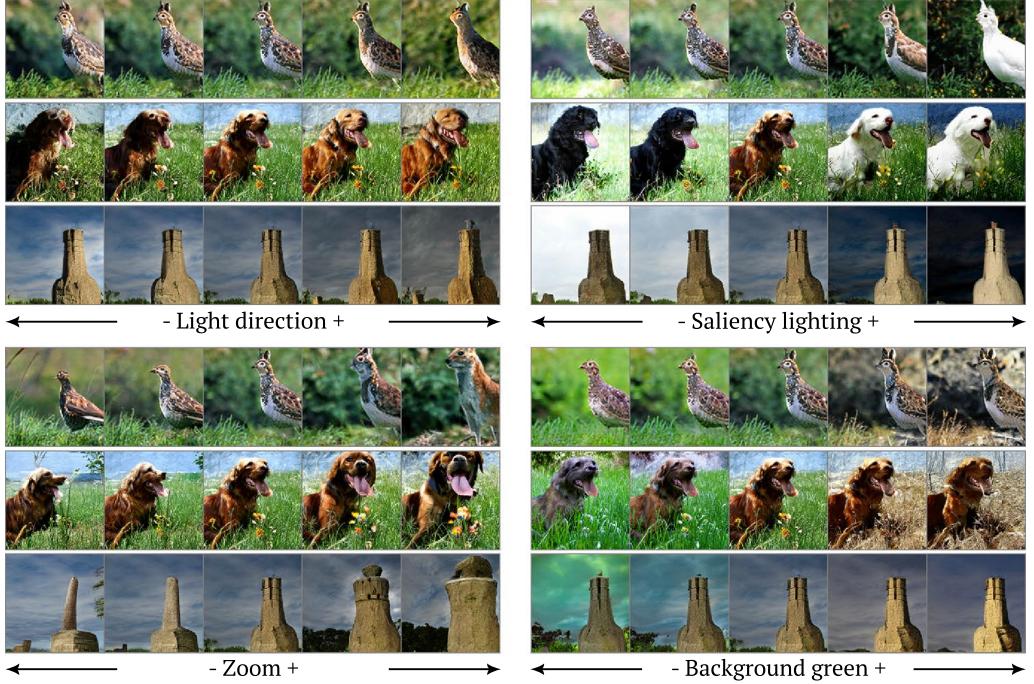


Figure 2: *Top*: images  $G(z)$ ; *Middle*: images after a latent shift  $G(z + h_{bg})$ ; *Bottom*: saliency masks

the latent representations  $\{E(I_1), \dots, E(I_N)\} \subset \mathbb{R}^{120}$  and sample the codes from the neighborhood of these representations. Formally, the samples have the form  $\{E(I_i) + \alpha \xi \mid i \sim \mathcal{U}\{1, N\}, \xi \sim \mathcal{N}(0, I)\}$ . Here  $\alpha$  denotes the neighborhood size and it should be larger for small  $I$  to prevent overfitting. In particular, we use  $\alpha=0$  for Imagenet and  $\alpha=0.2$  for all other cases.

**Mask size filtering.** Since some of the BigBiGAN-produced images are low-quality and do not contain clear objects, the corresponding masks can result in a very noisy supervision. To avoid this, we apply a simple filtering that excludes the images where the ratio of foreground pixels exceeds 0.5.

**Histogram filtering.** Since  $G(z+h_{bg})$  should have mostly dark and light pixels, we filter out the images that are not contrastive enough. Formally, we compute the intensity histogram with 12 bins for the grayscaled  $G(z+h_{bg})$ . Then we smooth it by taking the moving average with a window 3 and filter out the samples that have local maxima outside the first/last buckets of the histogram.

**Connected components filtering.** For each generated mask  $M$  we group the foreground pixels into connected (by edges) groups forming clusters  $M_1, \dots, M_k$ . Assuming that  $M_1$  is the cluster with the maximal area, we exclude all the clusters  $M_i$  with  $|M_i| < 0.2 \cdot |M_1|$ . This technique allows to remove visual artifacts from the synthetic data.

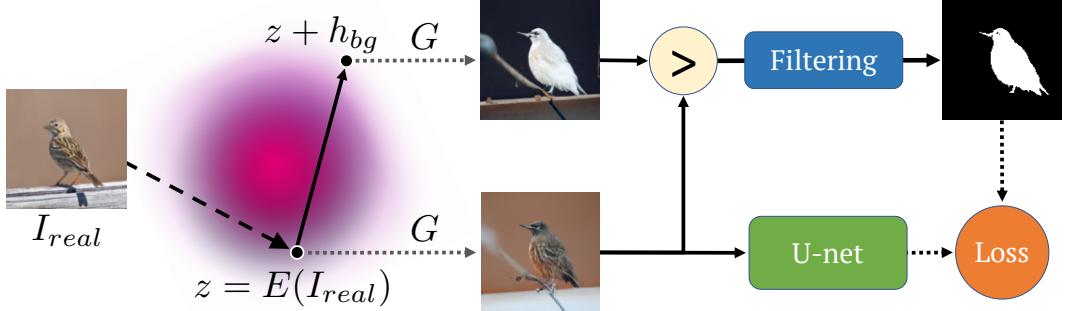


Figure 3: Examples of mask improvement. *Left*: sample rejected by the mask size filter. *Middle*: sample rejected by the histogram filtering. *Right block*: mask pixels removed by the connected components filter are shown in blue and the remaining mask pixels are shown in red.

### 3.3 Training model on synthetic data

Given a large amount of synthetic data, one can train one of the existing image-to-image CNN architectures in the fully-supervised regime. The whole pipeline is schematically presented in Figure 4. In all our experiments we employ a simple U-net architecture [12]. We train U-net on the synthetic dataset with Adam optimizer and the binary cross-entropy objective applied on the pixel level. We perform  $12 \cdot 10^3$  steps with batch 95. The initial learning rate equals 0.001 and is decreased by 0.2 on step  $8 \cdot 10^3$ . During inference, we rescale an input image to have a size 128 along its shorter side and scale the color channels to  $[-1, 1]$ . Compared to existing unsupervised alternatives, the training of our model is extremely simple, does not include a large number of hyperparameters. The only hyperparameters in our protocol are batch size, learning rate schedule, and a number of optimizer steps and we tune them on the hold-out validation set of synthetic data. Training with on-line synthetic data generation takes approximately seven hours on two Nvidia 1080Ti cards.

Figure 4: Schematic representation of our approach.



## 4 Experiments

The goal of this section is to confirm that the usage of GAN-produced synthetic data is a promising direction for unsupervised saliency detection and object segmentation. To this end, we extensively compare our approach to the existing unsupervised counterparts on the standard benchmarks.

**Evaluation metrics.** All the methods are compared in terms of the three measures described below.

- **F-measure** is an established measure in the saliency detection literature. It is defined as  $F_\beta = \frac{(1+\beta^2)\text{Precision} \times \text{Recall}}{\beta^2\text{Precision} + \text{Recall}}$ . Here Precision and Recall are calculated based on the binarized predicted masks and groundtruth masks as  $\text{Precision} = \frac{TP}{TP+FP}$  and  $\text{Recall} = \frac{TP}{TP+FN}$ , where TP, TN, FP, FN denote true-positive, true-negative, false-positive, and false-negative, respectively. We compute F-measure for 255 uniformly distributed binarization thresholds and report its maximum value  $\max F_\beta$ . We use  $\beta=0.3$  for consistency with existing works.
- **IoU** (intersection over union) is calculated on the binarized predicted masks and groundtruth as  $\text{IoU}(s, m) = \frac{\mu(s \cap m)}{\mu(s \cup m)}$ , where  $\mu$  denotes the area. The binarization threshold is set to 0.5.
- **Accuracy** measures the proportion of pixels that have been correctly assigned to the object/background. The binarization threshold for masks is set to 0.5.

Since the existing literature uses different benchmark datasets for saliency detection and object segmentation, we perform a separate comparison for each task below.

#### 4.1 Object segmentation.

**Datasets.** We use two following datasets from the literature of segmentation with generative models.

- **Caltech-UCSD Birds 200-2011** [28] contains 11788 photographs of birds with segmentation masks. We follow [7], and use 10000 images for our training subset and 1000 for the test subset from splits provided by [7]. Unlike [7], we do not use any images for validation and simply omit the remaining 788 images.
- **Flowers** [29] contains 8189 images of flowers equipped with saliency masks generated automatically via the method developed for flowers. In experiments with the Flowers dataset, we do not apply the mask area filter in our method, since it rejects most of the samples.

On these two datasets we compare the following methods:

- **PerturbGAN**[8] segments an image based on the idea that object location can be perturbed without affecting the scene realism. For comparison, we use the numbers reported in [8].
- **ReDO**[7] produces segmentation masks based on the idea that object appearance can be changed without affecting image quality. For comparison, we report the numbers from [7].
- **BigBiGAN** is our method where the latent codes are sampled from  $z \sim \mathcal{N}(0, \mathbb{I})$ .
- **E-BigBiGAN (w/o  $z$ -noising)** is our method where the latent codes of synthetic data are sampled from the outputs of the encoder  $E$  applied to the train images of the dataset at hand.
- **E-BigBiGAN (with  $z$ -noising)** same as above with latent codes sampled from the vicinity of the embeddings with the neighborhood size  $\alpha$  set to 0.2.

The comparison results are provided in Table 1, which demonstrates the significant advantage of our scheme. Note, since, both datasets in this comparison are small-scale,  $z$ -noising considerably improves the performance, increasing diversity of training images.

Method	CUB-200-2011			Flowers		
	max $F_\beta$	IoU	Accuracy	max $F_\beta$	IoU	Accuracy
PerturbGAN	—	0.380	—	—	—	—
ReDO	—	0.426	0.845	—	0.764	0.879
BigBiGAN	0.794	0.683	0.930	0.760	0.540	0.765
E-BigBiGAN (w/o $z$ -noising)	0.750	0.619	0.918	0.814	0.689	0.874
E-BigBiGAN (with $z$ -noising)	<b>0.834</b>	<b>0.710</b>	<b>0.940</b>	<b>0.878</b>	<b>0.804</b>	<b>0.904</b>
std	0.005	0.007	0.002	0.001	<0.001	<0.001

Table 1: The comparison of unsupervised object segmentation methods. For our model we report the performance averaged over ten runs. For the best model we also report the standard deviation values.

#### 4.2 Saliency detection.

**Datasets.** We use the following established benchmarks for saliency detection. For all the datasets groundtruth pixel-level saliency masks are available.

- **ECSSD**[30] contains 1,000 images with structurally complex natural contents.
- **DUTS**[31] contains 10,553 train and 5,019 test images. The train images are selected from the ImageNet detection train/val set. The test images are selected from the ImageNet test and the SUN dataset[32]. We always report the performance on the DUTS-test subset.
- **DUT-OMRON**[19] contains 5,168 images of high content variety.

**Baselines.** While there are a large number of papers on unsupervised deep saliency detection, all of them employ pretrained supervised models in their training protocols. Therefore, we use the

most recent “shallow” methods HS[33], wCtr[34], and WSC[35] as the baselines. These three methods were chosen based on their state-of-the-art performance reported in the literature and publicly available implementations. The results of the comparison are reported in Table 2. In this table, BigBiGAN denotes the version of our method where the latent codes of synthetic images are sampled from  $z \sim \mathcal{N}(0, \mathbb{I})$ . In turn, in E-BigBiGAN,  $z$  are sampled from the latent codes of Imagenet-train images, for all three datasets. Since the Imagenet dataset is large enough, we do not employ  $z$ -noising in this comparison.

Method	ECSSD			DUTS			DUT-OMRON		
	$\max F_\beta$	IoU	Accuracy	$\max F_\beta$	IoU	Accuracy	$\max F_\beta$	IoU	Accuracy
HS	0.673	0.508	0.847	0.504	0.369	0.826	0.561	0.433	0.843
wCtr	0.684	0.517	0.862	0.522	0.392	0.835	0.541	0.416	0.838
WSC	0.683	0.498	0.852	0.528	0.384	0.862	0.523	0.387	<b>0.865</b>
BigBiGAN	0.782	0.672	0.899	0.608	0.498	0.878	0.549	0.453	0.856
E-BigBiGAN	<b>0.797</b>	<b>0.684</b>	<b>0.906</b>	<b>0.624</b>	<b>0.511</b>	<b>0.882</b>	<b>0.563</b>	<b>0.464</b>	0.860

Table 2: The comparison of unsupervised saliency detection methods. For BigBiGAN and E-BigBiGAN we report the mean values over 10 independent runs.

As one can see, our method mostly outperforms the competitors by a considerable margin, which confirms the promise of using synthetic imagery in the unsupervised scenario. Several qualitative segmentation samples are provided on Figure 5.

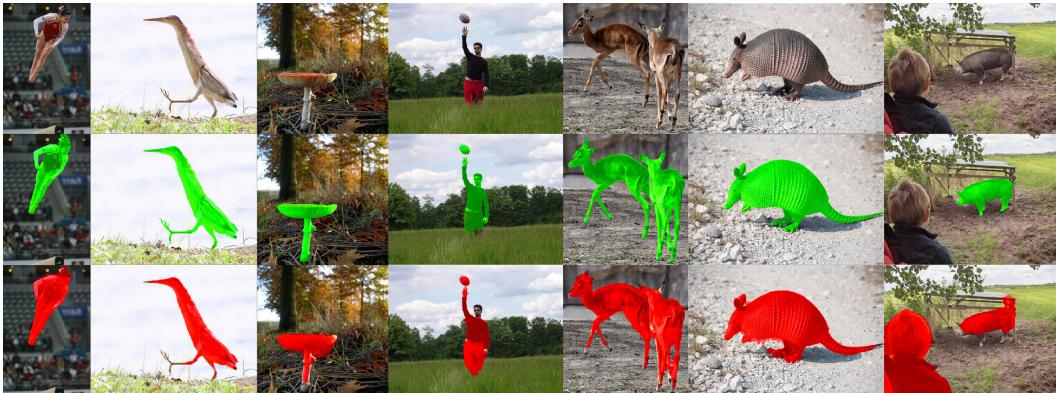


Figure 5: *Top*: Images from the DUTS-test dataset. *Middle*: Groundtruth masks. *Bottom*: Masks produced by the E-BigBiGAN method.

### 4.3 Weakly-supervised object localization (WSOL)

A closely related to segmentation problem is object localization, where for a given image one has to provide a bounding box instead of a segmentation mask. In this section, we demonstrate that our unsupervised method performs on par with the weakly-supervised state-of-the-art. To compare with the previous literature, we use the numbers from the very recent evaluation paper [2] that reviews a large number of existing WSOL methods and reports actual state-of-the-art. We employ exactly the same evaluation protocols as in [2] and compare the prior works with our E-BigBiGAN method, which samples  $z$  from the latent codes of Imagenet-train images, as described in Section 3.2. The comparison results are provided in Table 3.

**Evaluation metrics.** For the WSOL problem we use the following metrics [2]:

- **MaxBoxAcc** [36, 1]. For an image  $I_n$ , let us have a predicted mask  $s_n$  and a set of ground truth bounding boxes  $B_n^{(i)}$  for  $i = 1, \dots, m$  (some datasets can provide several bounding boxes per image). Let us select a threshold  $\tau \in [0, 1]$  and denote  $c_n^\tau$  the largest (in terms of the area) connected component of the mask  $s_n$  binarized with threshold  $\tau$ . Let us denote

with  $\text{box}(c_n^\tau)$  the minimal bounding box containing the set  $c_n^\tau$ . Then we define

$$\text{BoxAcc}(\tau) = \frac{1}{N} \sum_{n=1}^N \mathbb{1}_{\text{IoU}(\text{box}(c_n^\tau), B_n^{(j)}) \geq 0.5} \quad (1)$$

where  $B_n^{(j)}$  corresponds to the ground truth bounding box with the maximal IoU with  $\text{box}(c_n^\tau)$  and  $N$  denotes the number of images. Then the final metrics MaxBoxAcc is the maximum of  $\text{BoxAcc}(\tau)$  over all thresholds  $\tau$ .

- **PxAP** [37]. Let us have a predicted mask  $s_n$  and ground truth mask  $t_n$ . For a threshold  $\tau \in [0, 1]$  we define a pixel precision and recall

$$P_\tau = \frac{1}{N} \sum_{n=1}^N \frac{|\{s_n \geq \tau\} \cap \{t_n = 1\}|}{|\{s_n \geq \tau\}|}; \quad R_\tau = \frac{1}{N} \sum_{n=1}^N \frac{|\{s_n \geq \tau\} \cap \{t_n = 1\}|}{|\{t_n = 1\}|} \quad (2)$$

We average both values over all images and then PxAP is defined as the area under curve of the pixel precision-recall curve.

**Datasets.** We use the following benchmarks for weakly-supervised object localization.

- **Imagenet** [36]. For evaluation we use 10,000 validation images. The dataset contains several annotated bounding boxes for each image.
- **Caltech-UCSD Birds 200-2011** [28]. For evaluation we use 5,794 test images.
- **OpenImages** [2] contains a subset of OpenImages instance segmentation dataset [38]. For evaluation we use 5,000 randomly selected images from 100 classes as in [2].

Method	Imagenet (MaxBoxAcc)	CUB (MaxBoxAcc)	OpenImages (PxAP)
Previous SOTA [2]	0.654	0.781	0.630
E-BigBiGAN	0.614	0.742	0.638

Table 3: The comparison of E-BigBiGAN to the WSOL state-of-the-art. For E-BigBiGAN we report the mean values over 10 independent runs. Despite being completely unsupervised, E-BigBiGAN performs on par with the WSOL methods, which were trained under more supervision.

#### 4.4 Ablation.

In Table 4 we demonstrate the impacts of individual components in our method. First, we start with a saliency detection model trained on the synthetic data pairs  $\{G(z), M = [G(z+h_{bg}) > G(z)]\}$  with  $z \sim \mathcal{N}(0, I)$ . Then we add one by one the components listed in Section 3.2. The most significant performance impact comes from using the latent codes of the real images from the Imagenet.

Method	ECSSD			DUTS			DUT-OMRON		
	$\max F_\beta$	IoU	Accuracy	$\max F_\beta$	IoU	Accuracy	$\max F_\beta$	IoU	Accuracy
Base	0.737	0.626	0.859	0.575	0.454	0.817	0.498	0.389	0.758
+Imagenet embeddings	0.773	0.657	0.874	0.616	0.483	0.832	0.533	0.413	0.772
+Size filter	0.781	0.670	0.900	0.62	0.499	0.871	0.552	0.443	0.842
+Histogram	0.779	0.670	0.900	0.621	0.503	0.875	0.555	0.450	0.850
+Connected components	<b>0.797</b>	<b>0.684</b>	<b>0.906</b>	<b>0.624</b>	<b>0.511</b>	<b>0.882</b>	<b>0.563</b>	<b>0.464</b>	<b>0.860</b>

Table 4: Impact of different components in the E-BigBiGAN pipeline.

## 5 Conclusion

In our paper, we continue the line of works on unsupervised object segmentation with the aid of generative models. While the existing unsupervised techniques require adversarial training, we

introduce an alternative research direction, based on the high-quality synthetic data from the off-the-shelf GAN. Namely, we utilize the images produced by the BigBiGAN model, which is trained on the Imagenet dataset. Exploring BigBiGAN, we have discovered that its latent space semantics allows to produce the saliency masks for synthetic images automatically via latent space manipulations. As shown in experiments, this synthetic data is an excellent source of supervision for discriminative computer vision models. The main feature of our approach is its simplicity and reproducibility since our model does not rely on a large number of components/hyperparameters. On several common benchmarks, we demonstrate that our method achieves superior performance compared to existing unsupervised competitors.

We also highlight the fact that the state-of-the-art generative models, such as BigBiGAN, can be successfully used to generate training data for yet another computer vision task. We expect that other problems such as semantic segmentation can also benefit from the usage of GAN-produced data in the weakly-supervised or few-shot regimes. Since the quality of GANs will likely improve in the future, we expect that the usage of synthetic data will become increasingly widespread.

## References

- [1] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.
- [2] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020.
- [3] Wenguan Wang, Qixia Lai, Huazhu Fu, Jianbing Shen, and Haibin Ling. Salient object detection in the deep learning era: An in-depth survey. *arXiv preprint arXiv:1904.09146*, 2019.
- [4] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.
- [5] Xide Xia and Brian Kulis. W-net: A deep model for fully unsupervised image segmentation. *arXiv preprint arXiv:1711.08506*, 2017.
- [6] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874, 2019.
- [7] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In *Advances in Neural Information Processing Systems*, pages 12705–12716, 2019.
- [8] Adam Bielski and Paolo Favaro. Emergence of object segmentation in perturbed generative models. In *Advances in Neural Information Processing Systems*, pages 7254–7264, 2019.
- [9] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, pages 10541–10551, 2019.
- [10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- [11] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. *arXiv preprint arXiv:2002.03754*, 2020.
- [12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [13] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2814–2821, 2014.
- [14] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2083–2090, 2013.

- [15] Houwen Peng, Bing Li, Haibin Ling, Weiming Hu, Weihua Xiong, and Stephen J Maybank. Salient object detection via structured matrix decomposition. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):818–832, 2016.
- [16] Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Chumping Hou. Co-saliency detection for rgbd images based on multi-constraint feature matching and cross label propagation. *IEEE Transactions on Image Processing*, 27(2):568–579, 2017.
- [17] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2014.
- [18] Y Wei, F Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *IEEE, ICCV*, 2012.
- [19] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013.
- [20] Wenguan Wang, Jianbing Shen, Ling Shao, and Fatih Porikli. Correspondence driven saliency transfer. *IEEE Transactions on Image Processing*, 25(11):5025–5034, 2016.
- [21] Fang Guo, Wenguan Wang, Jianbing Shen, Ling Shao, Jian Yang, Dacheng Tao, and Yuan Yan Tang. Video saliency detection using object proposals. *IEEE transactions on cybernetics*, 48(11):3159–3170, 2017.
- [22] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.
- [23] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017.
- [24] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9029–9038, 2018.
- [25] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4048–4056, 2017.
- [26] Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadil, Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepups: Deep robust unsupervised saliency prediction via self-supervision. In *Advances in Neural Information Processing Systems*, pages 204–214, 2019.
- [27] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- [28] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset, 2011.
- [29] Maria-Elena Nilsback and Andrew Zisserman. Delving into the whorl of flower segmentation. In *BMVC*, volume 2007, pages 1–10, 2007.
- [30] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2015.
- [31] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017.
- [32] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE, 2010.
- [33] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1155–1162, 2013.

- [34] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2814–2821, 2014.
- [35] Nianyi Li, Bilin Sun, and Jingyi Yu. A weighted sparse coding framework for saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5216–5223, 2015.
- [36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
- [37] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1597–1604. IEEE, 2009.
- [38] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11700–11709, 2019.