Information-Theoretic Segmentation by Inpainting Error Maximization

Pedro Savarese  
TTI-Chicago  
savarese@ttic.edu

Sunnie S. Y. Kim  
Princeton University  
suhk@cs.princeton.edu

Michael Maire  
University of Chicago  
mmaire@uchicago.edu

Greg Shakhnarovich  
TTI-Chicago  
greg@ttic.edu

David McAllester  
TTI-Chicago  
mcallester@ttic.edu

Abstract

We study image segmentation from an information-theoretic perspective, proposing a novel adversarial method that performs unsupervised segmentation by partitioning images into maximally independent sets. More specifically, we group image pixels into foreground and background, with the goal of minimizing predictability of one set from the other. An easily computed loss drives a greedy search process to maximize inpainting error over these partitions. Our method does not involve training deep networks, is computationally cheap, class-agnostic, and even applicable in isolation to a single unlabeled image. Experiments demonstrate that it achieves a new state-of-the-art in unsupervised segmentation quality, while being substantially faster and more general than competing approaches.¹

1. Introduction

Deep neural networks have significantly advanced a wide range of computer vision capabilities, including image classification [38, 55, 56, 27], object detection [22, 50, 40], and semantic segmentation [8, 73]. Nonetheless, neural networks typically require massive amounts of manually labeled training data to achieve state-of-the-art performance. Applicability to problems in which labeled data is scarce or expensive to obtain often depends upon the ability to transfer learned representations from related domains.

These limitations have sparked exploration of self-supervised methods for representation learning, where an automatically-derived proxy task guides deep network training. Subsequent supervised fine-tuning on a small labeled dataset adapts the network to the actual task of interest. A common approach to defining proxy tasks involves predicting one part of the data from another, e.g., geometric relationships [18, 45, 21], colorization [39], inpainting [49]. A recent series of advances focuses on learning representations through a contrastive objective [6, 65, 59, 57], and efficiency scaling such systems [25, 11, 24] to achieve parity with supervised pre-training.

Another class of approaches frames unsupervised learning within a generative modeling context, building upon, e.g., generative adversarial networks (GANs) [23] or variational autoencoders (VAEs) [36]. Donahue et al. [19, 20] formulate representation learning using a bidirectional GAN. Deep InfoMax [16] drives unsupervised learning by maximizing mutual information between encoder inputs and outputs. InfoGAN [12], which adds a mutual information maximization objective to a GAN, demonstrates that deep networks can learn to perform image classification without any supervision—at least for small-scale datasets.

Inspired by this latter result, we focus on the question of whether more complex tasks, such as image segmentation,
can be solved in a purely unsupervised fashion, without reliance on any labeled data for training or fine-tuning. We address the classic task of generic, category-agnostic segmentation, which aims to partition any image into meaningful regions (e.g., foreground and background), without relying on knowledge of a predefined set of object classes.

Here we introduce **Inpainting Error Maximization (IEM)** as an approach to unsupervised segmentation. This approach is motivated by the intuition that a segmentation into objects minimizes the mutual information between the pixels in the segments, and hence makes inpainting of one segment given the others difficult. This gives a natural adversarial objective where a segmenter tries to maximize, while an inpainter tries to minimize, inpainting error. However, rather than adopt an adversarial training objective we find it more effective to fix a basic inpainter and directly maximize inpainting error through a form of gradient descent on the segmentation. Our version of IEM is learning-free and can be applied directly to any image in any domain. Figure 1 shows example results for foreground-background segmentation derived from our IEM method which is diagrammed in Figure 2.

We show that the segmentations produced by the learning-free IEM segmenter can be used as noisy training labels to train a deep segmentation network which further improves our segmentation quality. This bootstrapping does not utilize human generated labels and our system has no equivalent of fine-tuning.

While IEM has a natural adversarial nature, we avoid employing a GAN. This contrasts with recent GAN-based unsupervised segmentation approaches, such as ReDO [10] and PerturbGAN [5], which Section 2 reviews in detail. Experiments in Section 4 demonstrate that our learning-free method matches or outperforms both. Our work advances unsupervised segmentation via the following contributions:

- An information-theoretic inspired IEM procedure for image segmentation which is fast, learning-free, and can be applied directly to any image in any domain.
- Extensive empirical results showing that our IEM procedure performs competitively with prior work on unsupervised segmentation when measured in terms of intersection-over-union (IoU).
- An optional refinement phase for IEM wherein a neural segmentation network is trained on a subset of images and their IEM segmentations and where the training images are selected to be those having high IEM inpainting error. This network can then be incorporated into the IEM process, resulting in a system that comfortably outperforms all competing methods.

Our results put dominant approaches to unsupervised segmentation into question. In comparison to IEM, generative modelling not only results in more computationally expensive methods, but also fails at learning high-quality segmentations. IEM provides a new combination of modeling and learning, and perhaps a new direction for unsupervised methods.

### 2. Related Work

Semantic segmentation, as a pixel-level category labeling problem, has rapidly advanced due to availability of large-scale labeled datasets and supervised deep learning [42, 51, 2, 74, 7, 9]. These tools have yielded similarly impressive progress on segmentation of detected object instances [22, 26]. In absence of large-scale annotation, the now common self-supervised approach, consisting of proxy task driven representation learning followed by fine-tuning, has demonstrated successful transfer to semantic and object segmentation tasks [25]. Various efforts have also explored weak supervision from image labels [48, 47, 28], bounding boxes [66, 15, 47, 34], or saliency maps [46, 72, 62].

Less clear is the suitability of these weakly- or self-supervised approaches for category-agnostic image segmentation, which could be regarded as primarily a grouping or clustering problem, as opposed to a labeling task. Indeed, class-agnostic segmentation has inspired a diverse array of algorithmic approaches. One framework that has proven amenable to supervised deep learning is that of considering the dual problem of edge detection [4, 53, 67, 37] and utilizing classic clustering algorithms and morphological transformations to convert edges into regions [54, 1].

Unsupervised class-agnostic segmentation methods instead often directly address the partitioning task. Earlier work clusters pixels using color, contrast, and hand-crafted features [13, 32, 76]. More recent strategies include self-training a convolutional neural network [33], and maximizing mutual information between cluster assignments [31].

A particular line of recent unsupervised work uses generative models to partition an image. Chen et al. [10] propose a GAN-based object segmentation model, ReDO (Re-Drawing of Objects), premised on the idea that it should be possible to change the appearance of objects without affecting the realism of the image containing them. During training, the model’s mask generator infers multiple masks for a given image, and the region generator “redraws” or generates new pixel values for each mask’s region, one at a time. The discriminator then judges the newly composed image for realism. After training, passing a novel image through the learned mask generator suffices to segment it.

Bielski & Favaro [5] propose a different GAN-based model, which we refer to as PerturbGAN. They build on the idea that one can perturb object location without affecting image realism. Comprising their model is an encoder that maps an image into a latent code, a generator that constructs a layered representation consisting of a background image, a foreground image, and a mask, and a discriminator that assesses the layered representation. During training,
Figure 2. Inpainting Error Maximization (IEM) framework. Given an unlabeled image X, a mask generator module first produces segmentation masks (e.g., foreground M and background \( \bar{M} \)). Each mask selects a subset of pixels from the original image by performing an element-wise product between the mask and the image, hence partitioning the image into regions. Inpainting modules try to reconstruct each region given all others in the partition, and the IEM loss is defined by a weighted sum of inpainting errors.

small random shifts applied to the foreground object assist with learning. Inference proceeds by encoding the image and feeding the latent code to the trained generator.

In a different direction, Voynov et al. [60] examine the latent space of an off-the-shelf GAN and obtain saliency masks of synthetic images via latent space manipulations. Using these masks, they train a segmentation model with supervision. On another front, Benny & Wolf [3] train a complex model (OneGAN), with multiple encoders, generators, and discriminators, to solve several tasks simultaneously, including foreground segmentation. Their model is weakly-supervised, requiring class labels and clean background images but not pixel- or region-level annotation.

As modern techniques addressing the fully unsupervised setting, ReDO [10] and PerturbGAN [5] serve as a reference for experimental comparison to our IEM method. We do not compare to the methods of Voynov et al. [60] or Benny & Wolf [3], as the former involves manual examination of latent space manipulation directions and the latter utilizes weak supervision from additional datasets.

Lastly, our methodology relates to other recent information-theoretic segmentation approaches, though our setting and detailed framework differ. Yang et al. [68] segment objects in video by minimizing the mutual information between motion field partitions, which they approximate with an adversarial inpainting network. We likewise focus on inpainting objectives, but in a manner not anchored to trained adversaries and not reliant on video dynamics.

Wolf et al. [63] segment cells in microscopy images by minimizing a measure of information gain between partitions. They approximate this information gain measure using an inpainting network, and segment images of densely populated cells in a hierarchical fashion. While philosophically aligned in terms of objective, our optimization strategy and algorithm differs from that of Wolf et al. We search for a partition over the global image context, whereas Wolf et al. focus on partitioning local image patches in a manner reminiscent of edge detection.

3. Information-theoretic Segmentation

We model an image as a composition of independent layers, e.g., foreground and background, and aim to recover these layers by partitioning so as to maximize the error of inpainting one partition from the others. Our full method utilizes Inpainting Expectation Maximization (IEM) as the primary component in the first of a two-phase procedure. In this first phase, a greedy binary optimization algorithm produces, independently for each image, a segmentation mask assigning each pixel to a partition, with the goal to maximize mutual inpainting error. We approximate the inpainting error by a simple filtering procedure yielding an objective that is easy and cheap to compute.

We emphasize that there is no learning in this initial phase, meaning that segmentation masks produced for each image are independent of any other images, resulting in a computationally cheap and distribution-agnostic subroutine—i.e., it can be applied even to a single unlabeled image and, unlike prior work, does not require segmented objects to be of similar semantic classes.

When we do have a collection of images sharing semantics, we can apply the second phase, wherein we select masks from the first phase with the highest inpainting error (which we deem the most successful) and use them as label supervision to train a segmentation network. This phase
refines suboptimal masks from the first phase, while also enforcing semantic consistency between segmentations. Note that our second phase is optional, and although we observe empirical improvements from training a neural segmentation model, the quality of masks produced by optimizing IEM alone are comparable and often superior to ones learned by competing methods.

3.1. Segmenting by Maximizing Inpainting Error

Consider an image \( X \in \mathcal{X} = \mathbb{R}^{C \times H \times W} \) with \( C \) channels, height \( H \) and width \( W \). We assume that \( X \) is generated by a stochastic process: first, foreground and background images \( F \in \mathbb{R}^{C \times H \times W} \) and \( B \in \mathbb{R}^{C \times H \times W} \) are drawn independently from distributions \( \mathcal{D}_F \) and \( \mathcal{D}_B \), respectively. Next, a binary segmentation mask \( M(F, B) \in \mathcal{M} = \{0, 1\}^{1 \times H \times W} \) is deterministically produced by \( F \) and \( B \). Finally, the image is generated by composing the foreground and background as \( X = F \odot M + B \odot \overline{M} = F|M + B|_M \), where \( \overline{M} = 1 - M \), \( F|M = F \odot M \), \( B|_M = B \odot \overline{M} \), and \( \odot \) denotes the Hadamard product.

Moreover, we assume that the mapping \( M \) is injective and can be equivalently represented as \( M = M(X) \). That is, the mask can be fully recovered from the image \( X \).

Our framework relies on information-theoretic measures, particularly the mutual information between two random variables \( A \) and \( Z \):

\[
I(A, Z) = H(A) - H(A|Z) = H(Z) - H(Z|A),
\]

(1)

where \( H(A) \) is the entropy of \( A \) and \( H(A|Z) \) is the conditional entropy of \( A \) given \( Z \).

Mutual information between two random variables is zero if and only if they are independent. Under our assumption that \( F \) and \( B \) are independent, it follows that \( I(F, B) = 0 \). By the data-processing inequality of mutual information, it also follows that \( I(F|M, B|_M) = 0 \).

We consider the task of partitioning each image \( X \) into two sets of pixels \( \hat{F} \) and \( \hat{B} \) such that \( I(\hat{F}, \hat{B}) = 0 \). The hope is that \( \hat{F} = F|M \) and \( \hat{B} = B|_M \), hence fully recovering the ground-truth segmentation. However, this formulation admits a trivial solution where either \( \hat{F} \) or \( \hat{B} \) is empty, in which case we have \( I(\hat{F}, \hat{B}) = 0 \), a minimizer. We circumvent this issue by using a normalized variant of the mutual information, namely the coefficient of constraint [14]:

\[
C(A, Z) = \frac{I(A, Z)}{H(A)} + \frac{I(Z, A)}{H(Z)} = 2 - \left( \frac{H(A|Z)}{H(A)} + \frac{H(Z|A)}{H(Z)} \right).
\]

(2)

Similar to the mutual information, we have that \( C(A, Z) = 0 \) if and only if \( A \) and \( Z \) are independent. On the other hand, we have that one of the denominators approaches 0 as either \( A \) or \( Z \) approaches the empty set.

Therefore, partitioning each \( X \) into \( \hat{F} \) and \( \hat{B} \) to minimize \( C(\hat{F}, \hat{B}) \), as an alternative to minimizing \( I(\hat{F}, \hat{B}) \), partially avoids the trivial solution where either \( \hat{F} \) or \( \hat{B} \) is empty.

Now, let \( \phi : \mathcal{X} \rightarrow \mathcal{M} \) denote an arbitrary mapping from images to binary segmentation masks, and define random variables \( \hat{F}_\phi = X \odot \phi(X) \) and \( \hat{B}_\phi = X \odot \bar{\phi}(X) \), which are regions of the image partitioned by \( \phi \). Our goal is then to find \( \phi \) that minimizes the coefficient of constraint between the predicted foreground and background:

\[
\min_{\phi} C(\hat{F}_\phi, \hat{B}_\phi) = \max_{\phi} \frac{H(\hat{F}_\phi|\hat{B}_\phi)}{H(\hat{F}_\phi)} + \frac{H(\hat{B}_\phi|\hat{F}_\phi)}{H(\hat{B}_\phi)}. \tag{3}
\]

This optimization problem provides a general framework for unsupervised segmentation, where the distributions over \( \hat{F} \) and \( \hat{B} \) can be arbitrarily chosen.

We model the conditional pixel probabilities in Equation 3 as \( \ell_1 \)-Laplace densities with identity covariance matrices and conditional means \( \psi \), as these induce a \( \ell_1 \) reconstruction criterion which is commonly adopted in state-of-the-art image inpainting methods [71, 70].

For the marginals over \( \hat{F} \) and \( \hat{B} \) we opt for an agnostic approach which avoids introducing any bias on foreground and background pixel values. More specifically, we assign uniform distributions over pixel values, which amounts to
the approximations
\[ H(\hat{F}_\phi) \approx \mathbb{E}[\|\phi(X)\|] \quad H(\hat{B}_\phi) \approx \mathbb{E}[\|\bar{\phi}(X)\|]. \tag{4} \]

Appendix A discusses our choices of distributions in more detail, including how they lead to the approximations above. We also empirically compare our approximations against other options in Appendix C, adopting the same experimental protocol which is formalized later in Section 4.

Since the model \( \psi \) of the conditional means should be chosen such that the probability of observed foreground and background regions is maximized (i.e. the respective entropies are minimized), we must also account for optimization over \( \psi \). In this case, the optimization problem in Equation 3 becomes the following adversarial max–min problem, which we name Inpainting Error Maximization (IEM):
\[ \max_{\phi} \min_{\psi} \mathbb{E} \left[ \| X \circ \phi(X) - \psi(X \circ \bar{\phi}(X)) \|_1 \right] \]
\[ + \mathbb{E} \left[ \| X \circ \bar{\phi}(X) - \psi(X \circ \phi(X)) \|_1 \right]. \tag{5} \]

To illustrate the idea, we show qualitative results of inpainted foreground and background using ground-truth, eroded (smaller), and dilated (bigger) masks in Figure 3. We see that the ground-truth mask incurs high inpainting error for both the foreground and the background, while the eroded mask allows reasonable inpainting of the foreground and the dilated mask allows reasonable inpainting of the background. Hence we expect IEM, which maximizes the inpainting error of each partition given the others, to drive a predicted mask closer to the ground-truth.

3.2. Fast & Distribution-agnostic Segmentation

We design a procedure to approximately solve the IEM objective in Equation 5 given a single unlabeled image \( X \).

We start by introducing additional assumptions regarding our adopted image generation process: we assume that pixels of \( F \) and \( B \) have strong spatial correlation and hence can be approximately predicted by a weighted average of their neighbors. Let \( K \) denote the kernel of a Gaussian filter with standard deviation \( \sigma \) and arbitrary size, i.e.
\[ K_{i,j} \propto \frac{1}{2\pi\sigma^2} \exp \left( -\frac{(i-\mu_1)^2 + (j-\mu_2)^2}{2\sigma^2} \right), \tag{6} \]
where \( \mu_1, \mu_2 \) are the center pixel positions of the kernel, and \( K \) is normalized such that its elements sum to one. Then, we adopt the following inpainting module:
\[ \psi_K(X, M) = \frac{K \ast X}{K \ast M}. \tag{7} \]

Lastly, if the feasible set of \( \phi \) is arbitrarily expressive, e.g., the set of all functions from \( X' \) to \( M \), then optimizing \( \phi \) for a single image is equivalent to directly searching for a binary mask \( M \in \mathcal{M} = \{0,1\}^{1 \times H \times W} \):
\[ \max_{M \in \mathcal{M}} \frac{\| M \circ (X - \psi_K(X \circ \bar{M}, \bar{M})) \|_1}{\| M \|} \]
\[ + \frac{\| \bar{M} \circ (X - \psi_K(M \circ M, M)) \|_1}{\| \bar{M} \|}. \tag{8} \]

Let \( L_{\text{inp}} \) denote the above objective as a function of \( M \), for a fixed image \( X \). We optimize \( M \) with projected gradient ascent, adopting the update rule:
\[ M_{t+1} = \mathcal{P}_{S(M_t)} \left( M_t + \eta \nabla_M L_{\text{inp}}(M_t) \right), \tag{9} \]
where \( \eta \) is the learning rate, \( \mathcal{P} \) is the projection operator, and \( S(M_t) \subseteq \mathcal{M} \) is a local feasible set that depends on \( M_t \).

Our preliminary experiments show that taking \( \eta \to \infty \) results in faster increases of the inpainting objective, while also simplifying the method since, in this case, there is no need to tune the learning rate. The updates become:
\[ M_{t+1} = \mathcal{P}_{S(M_t)}^\infty \left( \nabla_M L_{\text{inp}}(M_t) \right), \tag{10} \]
where \( \mathcal{P}^\infty(M) = \lim_{\eta \to \infty} \mathcal{P}(\eta M) \), to be detailed below.

We define \( S(M) \) as the set of binary masks that differ from \( M \) only in its boundary pixels. More specifically, we say that a pixel is in the boundary of \( M \) if it contains a pixel in its 4-neighborhood with the opposite (binary) value. Therefore, the projection \( \mathcal{P}_{S(M)} \) has two roles: first, it restricts updates to pixels in the boundary of \( M \) (all other pixels of \( M \) have their values unchanged during updates); second, it ensures that new masks are necessarily binary by projecting each component of its argument to the closest value in \( \{0,1\} \). It then follows that \( \mathcal{P}_{S(M)}^\infty \) projects negative components to 0 and positive components to 1.

We also propose a regularization mechanism that penalizes the diversity in each of the extracted image layers. Let \( \sigma(M) \) denote the total deviation of pixels in \( X \circ M \):
\[ \sigma(M) = \| M \circ (X - \mu(M)) \|_2^2, \tag{11} \]
where \( \mu(M) \) is the average pixel value of \( X \circ M \) (not counting masked-out pixels). We define the regularized IEM objective as \( L_{\text{iem}}(M) = L_{\text{inp}}(M) - \frac{\lambda}{2} (\sigma(M) + \sigma(\bar{M})) \), where \( \lambda \) is a non-negative scalar that controls the strength of the diversity regularizer.

Finally, we adopt a simple smoothing procedure after each mask update to promote more uniform segmentation masks. For each pixel in \( M \), we compute the sum of pixel values in its 8-neighborhood, and assign a value of 1 to the pixel if the sum exceeds 4 and a value of 0 otherwise.
### 3.3. Mask Refinement by Iterative Learning

In this optional refinement phase, applicable when we have a dataset of unlabeled images known to share some semantics (e.g., same object category forming the foreground), we select masks produced by IEM with highest inpainting error (which we deem the most successful) and use them as labels to train a neural segmentation model with supervision. The goal is to improve mask quality and promote semantic consistency, as IEM performs distribution-agnostic segmentation. After the model is trained, we can, again optionally, return to the first phase and further refine the mask inferred by the segmentation model with IEM.

### 4. Experiments

We first describe our experimental setup and evaluation metrics. We then discuss qualitative results of IEM. Finally, we compare IEM quantitatively to recently proposed unsupervised segmentation methods, ReDO and PerturbGAN, and to the classic GrabCut algorithm.

#### 4.1. Datasets

We demonstrate IEM on the following datasets, and compare its performance to that of prior work where possible. **Caltech-UCSD Birds-200-2011 (CUB)** [61] consists of 11,788 images of 200 classes of birds and segmentation masks. **Flowers** [44] consists of 8,189 images of 102 classes of flowers, with segmentation masks obtained by an automated algorithm developed specifically for segmenting flowers in color photographs [43]. **LSUN Car** [69], as part of the large-scale LSUN dataset of 10 scene categories and 20 objects, consists of 5,520,753 car images. Segmentation masks are not provided, so following Bielski & Favaro [5], we approximate ground-truth masks for the first 10,000 images with Mask R-CNN [26], pre-trained on the COCO [41] dataset with a ResNet-50 FPN backend. We used the pre-trained model from the Detectron2 library [64]. Cars were detected in 9,121 images, and if several cars were detected, we grabbed the mask of the biggest instance as ground truth.

#### 4.2. Implementation Details

For all experiments, we adopt the same training, validation, and test splits as ReDO for CUB and Flowers (resulting in 1,000 and 1,020 test images, respectively), and random subsampling of 1,040 test images for LSUN Car. Preliminary experiments to guide the design of our method were performed on the validation set of CUB.

We train our segmentation model for a total of 50 epochs. Masks are initialized with centered squares of varying semantics (e.g., same object category forming the foreground) with a single Nvidia 1080 Ti. Running all 150 iterations of IEM on an entire set of ≈1000 images takes under 2 minutes on a single Nvidia 1080 Ti.

Masks are initialized with centered squares of varying sizes. For each dataset we run IEM, adopting squares with size 44, 78, and 92 as initialization, and only consider the results whose initialization lead to the highest inpainting error. Hence, ground truth labels are not used to choose mask initializations, and there is no feedback between test evaluation and the size of squares used to initialize the masks.

For the optional refinement phase, we select the 8,000 masks produced by IEM that induce the highest inpainting error and train a variant of PSPNet [74] to perform segmentation using the collected masks as pseudo-labels. The details of our segmentation model follow ReDO: it consists of an initial resolution-decreasing residual block, followed by three resolution-preserving residual blocks, and finally pyramid pooling, where all batch normalization layers [29] are replaced by instance norm [58].

We train our segmentation model for a total of 50 epochs.

### Table 1. Unsupervised segmentation results on Flowers, measured in terms of accuracy, IoU, mIoU, and DICE score. Segmentation masks used for evaluation are publicly available ground-truth.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>IoU</th>
<th>mIoU</th>
<th>DICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GrabCut [52]</td>
<td>82.0</td>
<td>69.2</td>
<td>70.0</td>
<td>79.1</td>
</tr>
<tr>
<td>ReDO [10]</td>
<td>87.9</td>
<td>76.4</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>IEM (ours)</td>
<td>88.5</td>
<td>76.8</td>
<td>79.0</td>
<td>84.6</td>
</tr>
<tr>
<td>IEM+SegNet (ours)</td>
<td><strong>89.6</strong></td>
<td><strong>78.9</strong></td>
<td><strong>80.8</strong></td>
<td><strong>86.0</strong></td>
</tr>
</tbody>
</table>

### Table 2. Unsupervised segmentation results on CUB, measured in terms of accuracy, IoU, mIoU, and DICE score. Segmentation masks used for evaluation are publicly available ground-truth.

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>IoU</th>
<th>mIoU</th>
<th>DICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GrabCut [52]</td>
<td>72.3</td>
<td>36.0</td>
<td>52.3</td>
<td>48.7</td>
</tr>
<tr>
<td>PerturbGAN [5]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>ReDO [10]</td>
<td>84.5</td>
<td>42.6</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>IEM (ours)</td>
<td>88.6</td>
<td>52.2</td>
<td>69.6</td>
<td>66.0</td>
</tr>
<tr>
<td>IEM+SegNet (ours)</td>
<td><strong>89.3</strong></td>
<td><strong>55.1</strong></td>
<td><strong>71.4</strong></td>
<td><strong>68.7</strong></td>
</tr>
</tbody>
</table>

### Table 3. Unsupervised segmentation results on LSUN Car, measured in terms of accuracy, IoU, mIoU, and DICE score. Segmentation masks used for evaluation were automatically generated with Mask R-CNN, following PerturbGAN [5].

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>IoU</th>
<th>mIoU</th>
<th>DICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GrabCut [52]</td>
<td>69.1</td>
<td>57.6</td>
<td>53.4</td>
<td>71.8</td>
</tr>
<tr>
<td>PerturbGAN [5]</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>IEM (ours)</td>
<td>76.2</td>
<td>65.1</td>
<td>61.5</td>
<td>78.1</td>
</tr>
<tr>
<td>IEM+SegNet (ours)</td>
<td><strong>77.8</strong></td>
<td><strong>68.5</strong></td>
<td><strong>63.2</strong></td>
<td><strong>80.5</strong></td>
</tr>
</tbody>
</table>
to minimize the pixel-wise binary cross-entropy loss. The network is optimized with Adam [35], a constant learning rate of $10^{-3}$, and a mini-batch size of 128.

4.3. Evaluation Metrics

In our framework, the predicted and ground-truth masks are binary and have 1 in foreground pixels and 0 in background pixels. We evaluate the predicted masks’ quality with four commonly used metrics. First, we measure the (per-pixel) mean accuracy of the foreground/background prediction. Second, we measure the predicted foreground region’s intersection over union (IoU) with the ground-truth foreground region. Third, we measure the mean IoU (mIoU) of the foreground and background regions. Finally, we measure the DICE score [17] defined as $\frac{2|\hat{F} \cap F|}{(|\hat{F}| + |F|)}$, where $\hat{F}$ is the predicted foreground region and $F$ is the ground-truth foreground region. In all metrics, higher values mean better performance.

4.4. Qualitative Analysis

In Figures 4, 5, 6, we show qualitative results of IEM; all the examples were sampled randomly from the respective test sets, without curation. In many cases, IEM accurately detects and segments the foreground object. It is successful for many flowers (Figure 4), where there is often a clear color difference between the foreground (flower) and the background (leaves), but also for many birds even when the background is more complex and the colors of background and foreground are similar. Moreover, we observe that the inpainting results are inversely correlated with the mask quality: better masks tend to produce worse inpainting, as expected under IEM. The optional refinement phase with the segmentation model (IEM+SegNet result) tends to improve the masks, albeit by a modest margin.

4.5. Quantitative Comparison to Prior Work

We next compare our method to two recent unsupervised segmentation methods, ReDO and PerturbGAN, and to the classic GrabCut algorithm that uses weak supervision from user-specified bounding-boxes. In Table 1, we compare segmentation results on the Flowers dataset. We find that masks from our first phase (IEM) outperform those from ReDO in terms of accuracy and IoU. We see further improvements for masks refined with the segmentation model (IEM+SegNet result). While accuracy and IoU are the only metrics reported for ReDO, we also report mIoU and the Dice score for completeness and future use. Overall, both the qualitative and the quantitative results suggest that the produced masks are high quality.

In Tables 2 and 3, we show results on CUB and LSUN Car. The recent GAN-based methods, ReDO and PerturbGAN, outperform GrabCut, but are likely limited by the known shortcomings of GANs such as mode collapse and unstable training. In contrast, IEM is computationally cheap, class-agnostic, and even applicable to a single unlabeled image. Consistent with Table 1, IEM outperforms GrabCut, ReDO, and PerturbGAN, often by a big margin. The refined masks (IEM+SegNet) show consistent further improvement, achieving superior performance across all metrics and datasets.

4.6. Ablation Studies

Studies regarding design choices of IEM are discussed in detail in the Appendix and briefly summarized here. Appendix B shows through ablation experiments that different components of IEM are necessary to achieve optimal results, namely regularizing the pixel-wise deviations (Equation 11), smoothing masks after each step, and restricting updates to the mask’s boundary.
Moreover, Appendix C investigates the distributional assumptions adopted for the IEM objective in Equation 5, showing that alternative approximations yield comparable, although inferior, segmentation results on CUB.

Finally, Appendix D studies the choice of the inpainting component in IEM, where experiments show that, perhaps surprisingly, adopting state-of-the-art inpainters causes considerable degradation of IEM’s segmentation performance.

5. Conclusion

IEM is a simple approach to unsupervised image segmentation that outperforms competitors reliant on far heavier learning machinery and more computationally expensive models. More broadly, our results demonstrate a successful recipe for unsupervised learning, distinct from both representation learning guided by proxy tasks, as well as deep learning-based generative modeling approaches.

This recipe calls for optimizing an objective strongly tied to the task of interest. Formulation of the optimization problem is actually agnostic to the use of learning. For unsupervised segmentation, we found a criterion motivated by information-theoretic considerations sufficiently powerful to advance the state-of-the-art. However, future instantiations of this recipe, applied to segmentation or other domains, could seek to bring learning machinery into this first phase. In either case, solutions to the optimization objective yield predictions for the target task, allowing a learning-centric second phase to distill and generalize these predictions by training a deep neural network to replicate them.

Acknowledgments: This work was in part supported by AFOSR Center of Excellence Award, FA9550-18-1-0166.
References


