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]

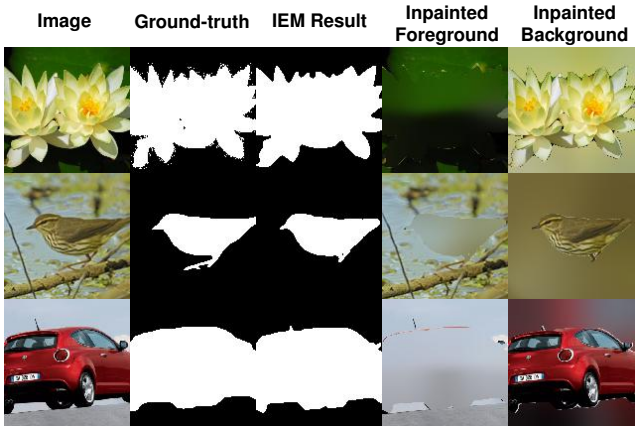

Figure 1. Illustration of our Inpainting Error Maximization (IEM) framework for completely unsupervised segmentation, applied to flowers, birds, and cars. Segmentation masks maximize the error of inpainting foreground given background and vice-versa.

1. Introduction

Deep neural networks have significantly advanced a wide range of computer vision capabilities, including image classification [38, 58, 59, 27], object detection [22, 53, 42], and semantic segmentation [8, 77]. 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, 47, 21], colorization [40], inpainting [51]. A recent series of advances focuses on learning representations through a contrastive objective [6, 68, 62, 60], 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,

[1] Code is available at https://github.com/lolemacs/iem

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. IEM 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 maxmize, 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 [44, 54, 2, 78, 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 [50, 49, 28], bounding boxes [69, 15, 49, 34], or saliency maps [48, 76, 65].

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, 56, 70, 37] and utilizing classic clustering algorithms and morphological transformations to convert edges into regions [57, 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, 80]. 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 (ReDrawing 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, small random shifts applied to the foreground object assist

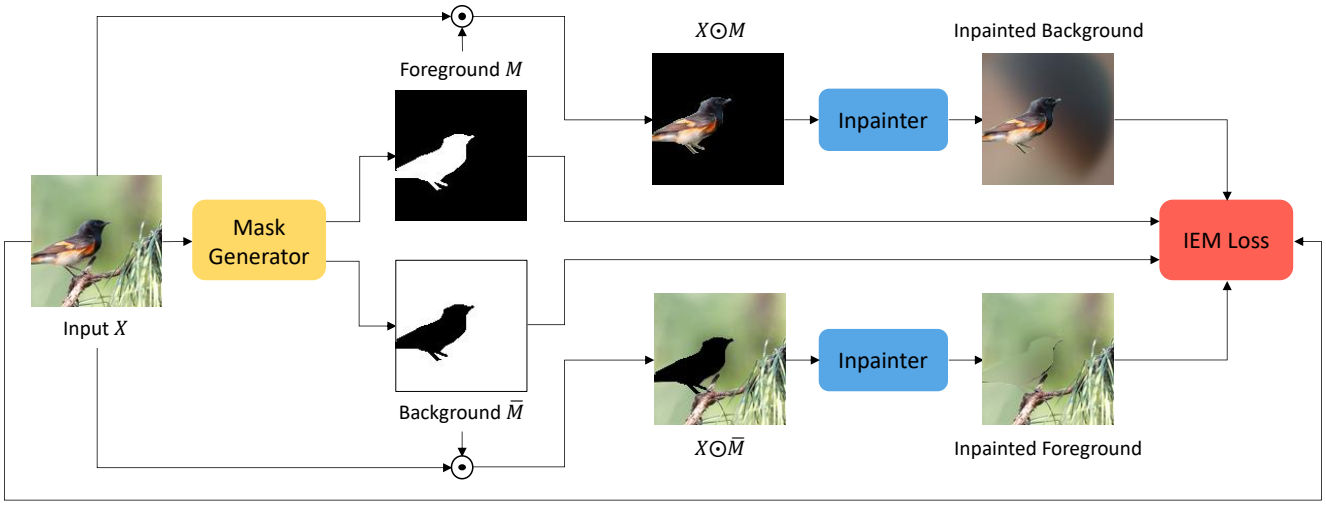

Figure 2. Inpainting Error Maximization (IEM) framework. Given an unlabeled image X, a mask generator module first produces segmentation masks (*e.g.*, foreground M and background $\bar{M}$). Each mask selects a subset of pixels from the original image by performing an element-wise product between the 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.

with learning. Inference proceeds by encoding the image and feeding the latent code to the trained generator.

In a different direction, Voynov *et al*. [63] 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*. [63] or Benny & Wolf [3], as the former involves manual examination of latent space manipulation directions and the latter utilizes weak supervision from additional datasets.

Our methodology relates to other recent information-theoretic segmentation approaches, though our setting and framework differ. Yang *et al*. [72, 71] segment objects in video by minimizing the mutual information between motion field partitions, which they approximate with an adversarial inpainter. We likewise focus on inpainting objectives, but in a manner not anchored to trained adversaries and not reliant on video dynamics.

Wolf *et al*. [66] 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 Error 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 en-

forcing 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:

$$I(A,Z) = H(A) - H(A|Z) = H(Z) - H(Z|A)\,, \quad (1)$$

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 problem 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, 39, 52]:

$$\begin{aligned} C(A,Z) &= \frac{I(A,Z)}{H(A)} + \frac{I(Z,A)}{H(Z)} \\ &= 2 - \left(\frac{H(A|Z)}{H(A)} + \frac{H(Z|A)}{H(Z)}\right). \end{aligned} \quad (2)$$

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

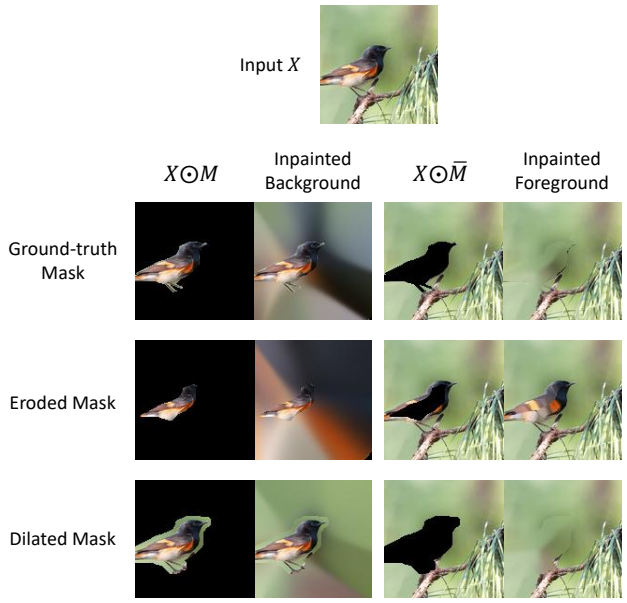

Figure 3. Foreground and background inpainting results using ground-truth, eroded (smaller), and dilated (bigger) masks. We see that the ground-truth mask incurs high inpainting error for both the foreground and the background, while the eroded mask allows reasonable inpainting of the foreground and 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 yield a segmentation mask close to the ground-truth.

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

$$\min_\phi \; C(\hat{F}_\phi, \hat{B}_\phi) = \max_\phi \; \frac{H(\hat{F}_\phi|\hat{B}_\phi)}{H(\hat{F}_\phi)} + \frac{H(\hat{B}_\phi|\hat{F}_\phi)}{H(\hat{B}_\phi)}. \quad (3)$$

This problem provides a general framework for unsupervised segmentation, where the distributions over $\hat{F}$ and $\hat{B}$ required to compute $C(\hat{F}_\phi, \hat{B}_\phi)$ can be arbitrarily chosen.

We model the conditional probabilities required to approximate $H(\hat{F}_\phi|\hat{B}_\phi)$ and $H(\hat{B}_\phi|\hat{F}_\phi)$ as ℓ_1-Laplacians with identity covariances and conditional means ψ, yielding expectations over $\|\hat{F}_\phi - \psi(\hat{B}_\phi)\|_1$ and $\|\hat{B}_\phi - \psi(\hat{F}_\phi)\|_1$.

For the marginals that define $H(\hat{F}_\phi)$ and $H(\hat{B}_\phi)$, 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 results in expectations over $\|\phi(X)\|$ and $\|\overline{\phi(X)}\|$.

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 ψ 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 ψ, which adds an adversarial component to the optimization problem in Equation 3:

$$\max_{\phi} \min_{\psi} \frac{\mathbb{E}\left[\|X \odot \phi(X) - \psi(X \odot \overline{\phi(X)})\|_1\right]}{\mathbb{E}\left[\|\phi(X)\|\right]} + \frac{\mathbb{E}\left[\|X \odot \overline{\phi(X)} - \psi(X \odot \phi(X)\|_1\right]}{\mathbb{E}\left[\|\overline{\phi(X)}\|\right]} . \tag{4}$$

Note that if the marginal and conditional probabilities required to characterize $C(\hat{F}_\phi, \hat{B}_\phi)$ in Equation 3 were instead chosen to be Gaussians, the objective would be equivalent to the Contextual Information Separation (CIS) criterion [72, 71] applied to raw images instead of motion fields. The objective above can thus be seen as a variant of CIS with different density models to approximate $C(\hat{F}_\phi, \hat{B}_\phi)$.

Objectives defined by an ℓ_1 inpainting error normalized by the mask size (*normalized inpainting error*), as in Equation 4, have also been employed to train inpainters in, for example, Yu *et al.* [74, 75][2] and Lin [41], giving another interpretation to the objective above: a maximization of the normalized inpainting error.

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 4 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 σ and arbitrary size, *i.e.*

$$K_{i,j} \propto \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(i-\mu_i)^2 + (j-\mu_j)^2}{2\sigma^2}\right) , \tag{5}$$

where μ_i, μ_j 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:

$$\psi_K(X, M) = \frac{K * X}{K * M} . \tag{6}$$

Lastly, if the feasible set of ϕ is arbitrarily expressive, *e.g.*, the set of all functions from $\mathcal{X}$ to $\mathcal{M}$, then optimizing ϕ for a single image is equivalent to directly searching for a binary mask $M \in \mathcal{M} = \{0, 1\}^{1 \times H \times W}$:

$$\max_{M \in \mathcal{M}} \frac{\|M \odot (X - \psi_K(X \odot \overline{M}, \overline{M}))\|_1}{\|M\|} + \frac{\|\overline{M} \odot (X - \psi_K(X \odot M, M))\|_1}{\|\overline{M}\|} \tag{7}$$

Let $\mathcal{L}_{inp}$ 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:

$$M_{t+1} = \mathcal{P}_{S(M_t)}\Big(M_t + \eta \nabla_M \mathcal{L}_{inp}(M_t)\Big) , \tag{8}$$

where η 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:

$$M_{t+1} = \mathcal{P}^{\infty}_{S(M_t)}\Big(\nabla_M \mathcal{L}_{inp}(M_t)\Big) , \tag{9}$$

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 $\mathcal{P}^{\infty}_{S(M)}$ 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 $X \odot M$:

$$\sigma(M) = \|M \odot (X - \mu(M))\|_2^2 , \tag{10}$$

where $\mu(M)$ is the average pixel value of $X \odot M$ (not counting masked-out pixels). We define the regularized IEM objective as $\mathcal{L}_{IEM}(M) = \mathcal{L}_{inp}(M) - \frac{\lambda}{2}\big(\sigma(M) + \sigma(\overline{M})\big)$, where λ is a non-negative scalar that controls the strength of the diversity regularizer.

[2] See version 1.0.0 of the official repository.

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 in case 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)** [64] consists of 11,788 images of 200 classes of birds and segmentation masks. **Flowers** [46] 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 [45]. **LSUN Car** [73], 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 [43] dataset with a ResNet-50 FPN backend. We used the pre-trained model from the Detectron2 library [67]. 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.

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

	Accuracy	IoU	DICE
GrabCut [55]	82.0	69.2	79.1
ReDO [10]	87.9	76.4	—
IEM (ours)	88.3	76.8	84.6
IEM+SegNet (ours)	**89.6**	**78.9**	**86.0**

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

	Accuracy	IoU	DICE
GrabCut [55]	72.3	36.0	48.7
PerturbGAN [5]	—	38.0	—
ReDO [10]	84.5	42.6	—
IEM (ours)	88.6	52.2	66.0
IEM+SegNet (ours)	**89.3**	**55.1**	**68.7**

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

	Accuracy	IoU	DICE
GrabCut [55]	69.1	57.6	71.8
PerturbGAN [5]	—	54.0	—
IEM (ours)	76.2	65.1	78.1
IEM+SegNet (ours)	**77.8**	**68.5**	**80.5**

We run IEM for a total of 150 iterations on the test set of each dataset, with an inpainter ψ_K whose convolutional kernel K is of size 21×21 and has a scale $\sigma = 5$, which, for computational reasons, we approximate by two stacked convolutions with kernel sizes 11×11. Moreover, we set the strength of the diversity regularizer as $\lambda = 0.001$.

All images are resized and center-cropped to 128×128 pixels. Running all 150 iterations of IEM on a set of $\approx$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 [78] to perform segmentation using the collected masks as pseudo-labels. Following ReDO, our model consists of an initial resolution-decreasing residual block, followed by three resolution-

preserving residual blocks and pyramid pooling, where all batch norm layers [29] are replaced by instance norm [61].

We train our segmentation model for a total of 50 epochs 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 three commonly used metrics. First, we measure the (per-pixel) mean *accuracy* of the foreground prediction. Second, we measure the predicted foreground region's *intersection over union (IoU)* with the ground-truth foreground region. Finally, we measure the *DICE score* [17] defined as $2|\hat{F} \cap F|/(|\hat{F}| + |F|)$, 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 inpanting, 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. 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). While accuracy and IoU are the only metrics reported for ReDO, we also report 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, outperfom 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.

Figure 4. Results of IEM on Flowers. *Left to right:* input image; ground truth mask; IEM mask; inpainting result, with every pixel inpainted as FG or BG according to the IEM mask; SegNet mask

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 10), smoothing masks after each step, and restricting updates to the mask's boundary.

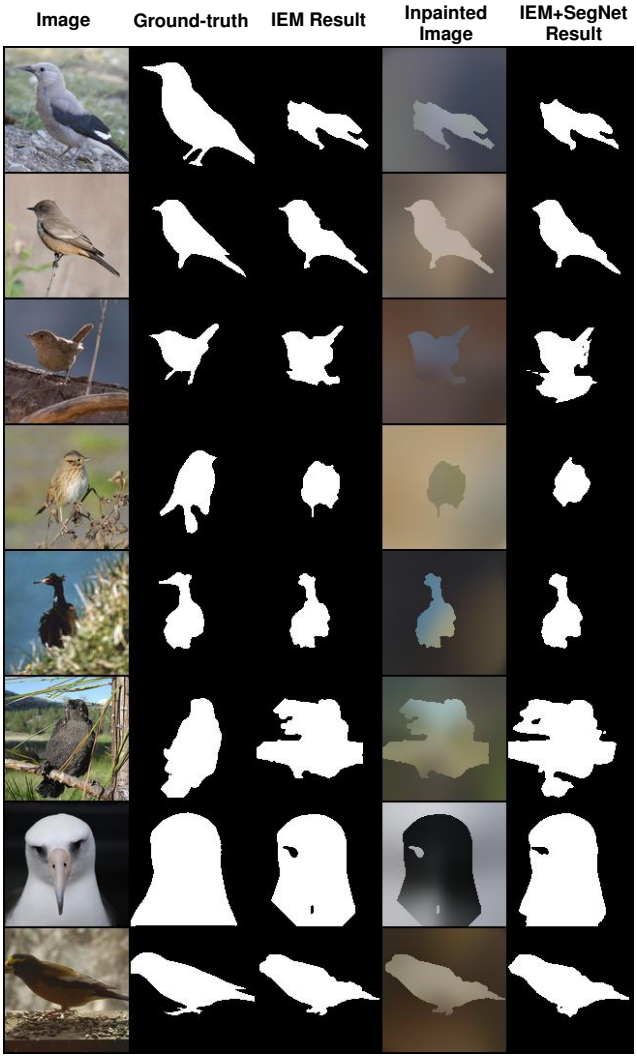

Figure 5. Results of IEM on CUB. *Left to right:* input image; ground truth mask; IEM mask; inpainting result, with every pixel inpainted as FG or BG according to the IEM mask; SegNet mask

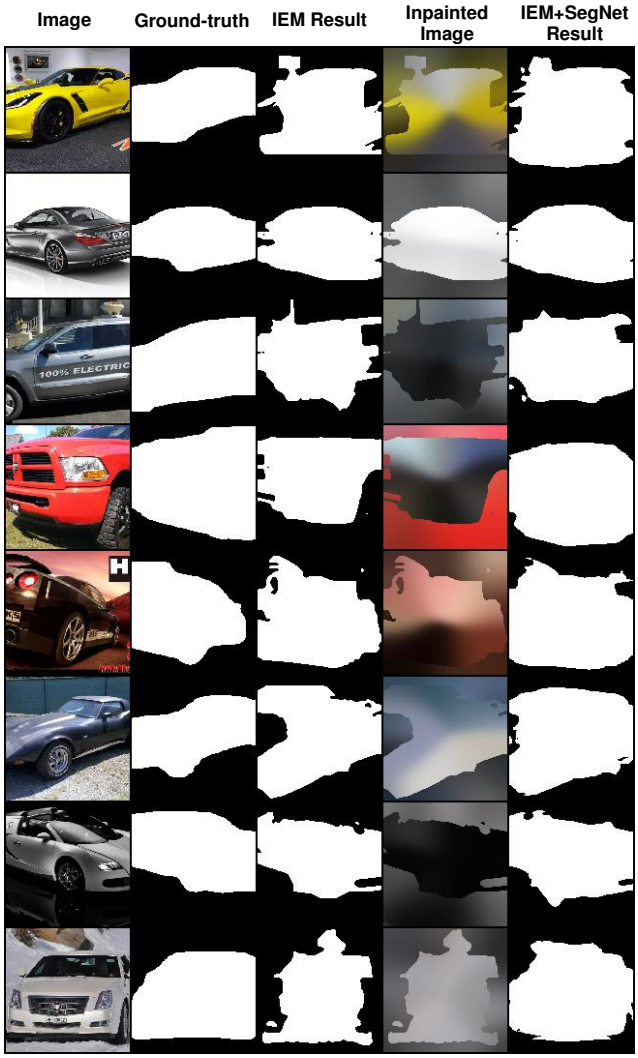

Figure 6. Results of IEM on LSUN Car. *Left to right:* input image; ground truth mask; IEM mask; inpainting result, with every pixel inpainted as FG or BG according to the IEM mask; SegNet mask

Moreover, Appendix C investigates the distributional assumptions adopted for the IEM objective in Equation 4, 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

[1] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. In *PAMI*, 2011.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. In *PAMI*, 2017.

[3] Yaniv Benny and Lior Wolf. OneGAN: Simultaneous unsupervised learning of conditional image generation, foreground segmentation, and fine-grained clustering. In *ECCV*, 2020.

[4] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani. Deepedge: A multi-scale bifurcated deep network for top-down contour detection. In *CVPR*, 2015.

[5] Adam Bielski and Paolo Favaro. Emergence of object segmentation in perturbed generative models. In *NeurIPS*, 2019.

[6] Piotr Bojanowski and Armand Joulin. Unsupervised learning by predicting noise. In *ICML*, 2017.

[7] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv:1706.05587*, 2017.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. In *PAMI*, 2017.

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[10] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In *NeurIPS*, 2019.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[12] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, 2016.

[13] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xialei Huang, and Shi-Min Hu. Global contrast based salient region detection. In *CVPR*, 2011.

[14] Clyde H. Coombs. *Mathematical psychology; an elementary introduction*. Prentice-Hall, 1970.

[15] Jifeng Dai, Kaiming He, and Jian Sun. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, 2015.

[16] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *ICLR*, 2019.

[17] Lee Raymond Dice. Measures of the amount of ecologic association between species. In *Ecology*, 1945.

[18] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[19] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2017.

[20] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *NeurIPS*, 2019.

[21] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[23] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. In *NeurIPS*, 2014.

[24] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[28] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, 2018.

[29] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[30] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[31] Xu Ji, Joao F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, 2019.

[32] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013.

[33] Asako Kanezaki. Unsupervised image segmentation by backpropagation. In *ICASSP*, 2018.

[34] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *CVPR*, 2017.

[35] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[36] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, 2014.

[37] Iasonas Kokkinos. Pushing the boundaries of boundary detection using deep learning. In *ICLR*, 2016.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *NeurIPS*, 2012.

[39] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. In *New Journal of Physics*, 2009.

[40] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *CVPR*, 2017.

[41] Hangyu Lin. https://github.com/avalonstrel/GatedConvolution_pytorch, 2018.

[42] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. In *ICCV*, 2017.

[43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

[44] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[45] Maria-Elena Nilsback and Andrew Zisserman. Delving into the whorl of flower segmentation. In *BMVC*, 2007.

[46] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[47] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.

[48] Seong Joon Oh, Rodrigo Benenson, Anna Khoreva, Zeynep Akata, Mario Fritz, and Bernt Schiele. Exploiting saliency for object segmentation from image level labels. In *CVPR*, 2017.

[49] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *ICCV*, 2015.

[50] Deepak Pathak, Philipp Krähenbühl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *ICCV*, 2015.

[51] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[52] William H. Press, Brian P. Flannery, and Saul A. Teukolsky. *Numerical recipes: The art of scientific computing*. Cambridge University Press, 1986.

[53] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *CVPR*, 2016.

[54] O. Ronneberger, P.Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

[55] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. GrabCut – interactive foreground extraction using iterated graph cuts. In *SIGGRAPH*, 2004.

[56] Wei Shen, Xinggang Wang, Yan Wang, Xiang Bai, and Zhijiang Zhang. Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *CVPR*, 2015.

[57] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. In *PAMI*, 2000.

[58] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[59] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[60] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv:1906.05849*, 2019.

[61] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv:1607.08022*, 2016.

[62] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

[63] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Big GANs are watching you: Towards unsupervised object segmentation with off-the-shelf generative models. *arXiv:2006.04988*, 2020.

[64] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[65] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *CVPR*, 2018.

[66] Steffen Wolf, Fred A. Hamprecht, and Jan Funke. Inpainting networks learn to separate cells in microscopy images. In *BMVC*, 2020.

[67] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[68] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. In *CVPR*, 2018.

[69] Wei Xia, Csaba Domokos, Jian Dong, Loong-Fah Cheong, and Shuicheng Yan. Semantic segmentation without annotating segments. In *ICCV*, 2013.

[70] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015.

[71] Yanchao Yang, Brian Lai, and Stefano Soatto. Time-supervised primary object segmentation. *arXiv:2008.07012*, 2020.

[72] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *CVPR*, 2019.

[73] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv:1506.03365*, 2015.

[74] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018.

[75] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019.

[76] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *ICCV*, 2019.

[77] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[78] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[79] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[80] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *CVPR*, 2014.

Appendix

A. Further Details on IEM Objective

The IEM objective in Equation 4 relies on adopting specific ℓ-norms to approximate the entropy terms in the coefficient of constraint. In particular, we rely on two main approximations, which we describe in detail below.

First, for the conditional entropy of the predicted foreground $\hat{F}_\phi$ given the predicted background $\hat{B}_\phi$ (and vice-versa), we have

$$\begin{aligned} H(\hat{F}_\phi|\hat{B}_\phi) &= H\left(X \odot \phi(X) \middle| X \odot \overline{\phi(X)}\right) \\ &= -\mathbb{E}_X\left[\log P\left(X \odot \phi(X) \middle| X \odot \overline{\phi(X)}\right)\right] \\ &\approx \mathbb{E}_X\left[\left\| X \odot \phi(X) - \psi(X \odot \overline{\phi(X)})\right\|_1\right], \end{aligned} \tag{11}$$

where the approximation adopted in the last step amounts to assigning a ℓ_1-Laplace distribution with identity covariance to the conditional pixel probabilities:

$$\begin{aligned} &P\left(X \odot \phi(X) \middle| X \odot \overline{\phi(X)}\right) \\ &\quad = \mathcal{L}\left(X \odot \phi(X); \mu\left(X \odot \overline{\phi(X)}\right), I\right) \\ &\quad \propto \exp\left(-\left\| X \odot \phi(X) - \mu\left(X \odot \overline{\phi(X)}\right)\right\|_1\right). \end{aligned} \tag{12}$$

Second, for the marginal entropies of the predicted foreground and background, we adopt

$$\begin{aligned} H(\hat{F}_\phi) &= H\left(X \odot \phi(X)\right) \\ &= -\mathbb{E}_X\left[\log P\left(X \odot \phi(X)\right)\right] \\ &\approx \mathbb{E}_X\left[\|\phi(X)\|\right], \end{aligned} \tag{13}$$

where $\|\phi(X)\|$ can be seen as any ℓ^p norm: since $\phi(X)$ is binary, we have that $\|\phi(X)\|_p = \|\phi(X)\|_q$ for any $p, q \in [1, \infty)$. Since modelling marginal distributions over images is known to be hard, we opt for an assumption-free approach and assume that pixel values are uniformly distributed, *i.e.* the approximation in the last step above corresponds to the assumption

$$P\left(X \odot \phi(X)\right) = \mathcal{U}(k)^{\|\phi(X)\|} = k^{-\|\phi(X)\|}, \tag{14}$$

where k captures the number of possible values for a pixel, *e.g.*, 255^3 for RGB images where each pixel channel is encoded as 8 bits. Note that $\|\phi(X)\|$ in the equation above represents the number of 1-valued elements in $\phi(X)$, hence it can be taken to be any ℓ_p norm (or any other function that matches this definition for binary inputs).

Table 4. Ablation experiments on CUB and Flowers. Number indicate IoU of masks produced by IEM.

	CUB	Flowers
Default parameters	52.2	76.8
No regularization on fore/back deviation	50.6	68.0
No smoothing on projection	47.0	75.7
Updates not restricted to mask boundary	42.8	76.6

Table 5. CUB results with different variants of the proposed IEM objective, each corresponding to different assigned distributions for conditional and marginal pixel distributions.

Objective (first term)	IoU	DICE
$\frac{\|M \odot (X - \psi_K(X \odot \overline{M}, \overline{M}))\|_1}{\|M\|}$ (Equation 7)	**52.2**	**66.0**
$\frac{\|M \odot (X - \psi_K(X \odot \overline{M}, \overline{M}))\|_2}{\|M\|}$ (Assumption 1)	51.7	65.6
$\frac{\|M \odot (X - \psi_K(X \odot \overline{M}, \overline{M}))\|_1}{\|X \odot M\|_1}$ (Assumption 3)	51.8	65.6
$\frac{\|M \odot (X - \psi_K(X \odot \overline{M}, \overline{M}))\|_2}{\|X \odot M\|_2}$ (Assumptions 1+2)	51.2	65.1

B. Analysis on Training Components of IEM

To understand the effect of IEM's components, we conduct ablation experiments on CUB and Flowers. We follow the same setup adopted for experiments in Section 4, running IEM for 150 iterations on the test set of each dataset.

First, we experiment with removing the regularization on foreground and background deviation (Equation 10) by setting $\lambda = 0$ in $\mathcal{L}_{IEM}$. Second, we remove the smoothing procedure after mask updates. Third, we allow mask updates at pixels other than the boundary.

In Table 4, we report IoU of produced masks for each experiment. Compared to the results with default parameters, mask quality drops in all three ablation experiments, suggesting that these components are important for IEM to achieve the best results. The regularization seems particularly important for Flowers, since it promotes homogeneous colors in the foreground and the background when the images have a clear color contrast between the two. Smoothing masks and limiting updates to the mask boundary seems more important in CUB, where the images have more complex backgrounds, as they prevent the bird segmentations from including other objects (*e.g.*, branches, grass).

C. Analysis on Approximations in IEM

As discussed in Appendix A, our proposed IEM objective adopts two key approximations for the conditional and marginal entropies in the original coefficient of constraint minimization problem in Equation 3. Although the Laplacian approximation for conditional pixel probabilities

is popular in the computer vision literature, for example in papers on inpainting [75, 74] and image modelling [30, 79], it is unclear whether it is the optimal choice for our setting.

Additionally, the uniform prior over pixel values that we adopt to approximate marginal entropies can be seen as being overly simple, especially since different priors are more commonly adopted in the literature *e.g.*, zero-mean isotropic Gaussians.

To investigate whether our approximations are sensible, we consider three variants of the proposed IEM objective, each being the result of different approximations for the image entropies. In particular, we consider:

1. Assuming that the conditional pixel probabilities follow a isotropic Gaussian (instead of a ℓ_1-Laplacian), which yields the approximation

$$H(\hat{F}_\phi|\hat{B}_\phi) \approx \mathbb{E}\left[\left\|X \odot \mu(X) - \psi(X \odot \overline{\phi(X)})\right\|_2\right], \quad (15)$$

   which in practice amounts to adopting the ℓ_2 norm instead of ℓ_1 in the numerators.

2. Assuming that the marginal foreground/background distributions are zero-mean isotropic Gaussians, which results in

$$H(\hat{F}_\phi) \approx \mathbb{E}\left[\|X \odot \phi(X)\|_2\right]. \quad (16)$$

3. Assuming that the marginal foreground/background distributions are zero-mean ℓ_1-Laplacians with identity covariance, yielding

$$H(\hat{F}_\phi) \approx \mathbb{E}\left[\|X \odot \phi(X)\|_1\right]. \quad (17)$$

We repeat our experiments on the CUB dataset, following the same protocol described in Section 4, *i.e.* masks are optimized for a total of 150 iterations to maximize the corresponding objective, and ψ_K is the same fixed inpainter as in our original experiments.

Table 5 summarizes our results, showing that although our chosen approximations yield the best segmentation performance measured in IoU and DICE score, all variants of the IEM objective offer comparable results. This suggests that our proposed framework does not strongly rely on our particular distributional assumptions (or, equivalently, to the adopted norms for the inpainting objective), offering a general approach for unsupervised segmentation.

D. Analysis on Inpainting Component

The inpainter we adopted for all experiments in Section 4 is significantly simpler than inpainting modules typically employed in other works, consisting of a single 21×21 convolution with a Gaussian filter. Such module has the advantage of having a small computational cost and not requiring any training, making it suitable for a learning-free method.

Table 6. Comparison between our simple inpainter and variants of the Gated Convolutional (GatedConv) model proposed in Yu *et al*. [75], in term of quality of masks produced on CUB. Removing components from GatedConv, such as removing its refinement phase during IEM ('GatedConv, coarse outputs') and training without adversarial losses ('GatedConv, ℓ_1 only') deteriorates its inpainting quality but results in better IEM segmentations.

Inpainting Module	IoU	DICE
Simple (Equation 6)	**52.2**	**66.0**
GatedConv [75]	40.3	55.8
GatedConv, coarse outputs [75]	41.6	56.8
GatedConv, ℓ_1 only [75]	43.7	59.0
GatedConv+Fine tuning, ℓ_1 only [75]	41.7	57.1

Here, we show that such simple inpainting module also yields better segmentation masks when compared to more sophisticated variants. Table 6 shows the quality of masks produced by IEM when adopting the inpainting component proposed in Yu *et al*. [75], which consists of gated convolutions and contextual attention, and is trained with the ℓ_1 loss along with an adversarial objective produced by a patch-wise discriminator ('GatedConv' entry in the table).

'GatedConv (coarse outputs)' refers to IEM results when taking the coarse outputs of GatedConv to compute the IEM objective: more specifically, we take the 'GatedConv' model (trained with both the ℓ_1 and adversarial loss) but only pass the foreground/background image through the first half of the network, which generates a coarse inpainted image that precedes the contextual attention layers (see Figure 3 of Yu *et al*. [75] for reference). 'GatedConv (ℓ_1 only)' refers to the GatedConv model trained only with the ℓ_1 loss (*i.e.* without SN-PatchGAN), with coarse outputs only. All GatedConv models were pre-trained on the whole CUB dataset with free-form masks [75] and then held fixed during IEM. When evaluated in terms of IoU and DICE, the quality of masks produced by IEM deteriorates monotonically with the complexity of the inpainting component: the original GatedConv model yields the lowest segmentation scores, which improves if IEM is run against its intermediate, coarse inpaintings, and training the network without contextual attention or the adversarial loss yields the best segmentation results other than ours.

We also evaluate how fine-tuning the inpainter during IEM, *i.e.* optimizing the inpainter with masks currently produced by IEM, affects the quality of segmentation masks. Table 6 shows that it also deteriorates the quality of masks produced by IEM (compare last two rows).