Appendix

A. Further Details on IEM Objective

The IEM objective in Equation 5 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

$$\begin{aligned} H(\hat{F}_\phi|\hat{B}_\phi) &= H\left(X \odot \phi(X) \middle| X \odot \overline{\phi(X)}\right) \\ &= -\mathbb{E}_X\left[\log P\left(X \odot \phi(X) \middle| X \odot \overline{\phi(X)}\right)\right] \\ &\approx \mathbb{E}_X\left[\left\| X \odot \phi(X) - \psi(X \odot \overline{\phi(X)})\right\|_1\right], \end{aligned} \tag{12}$$

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

$$\begin{aligned} &P\left(X \odot \phi(X) \middle| X \odot \overline{\phi(X)}\right) \\ &\quad = \mathcal{L}\left(X \odot \phi(X); \mu\left(X \odot \overline{\phi(X)}\right), I\right) \\ &\quad \propto \exp\left(-\left\| X \odot \phi(X) - \mu\left(X \odot \overline{\phi(X)}\right)\right\|_1\right). \end{aligned} \tag{13}$$

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

$$\begin{aligned} H(\hat{F}_\phi) &= H\left(X \odot \phi(X)\right) \\ &= -\mathbb{E}_X\left[\log P\left(X \odot \phi(X)\right)\right] \\ &\approx \mathbb{E}_X\left[\|\phi(X)\|\right], \end{aligned} \tag{14}$$

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

$$P\left(X \odot \phi(X)\right) = \mathcal{U}(k)^{\|\phi(X)\|} = k^{-\|\phi(X)\|}, \tag{15}$$

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 8)	**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 11) 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 [71, 70] and image modelling [30, 75], 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

$$H(\hat{F}_\phi|\hat{B}_\phi) \approx \mathbb{E}\left[\left\|X \odot \mu(X) - \psi(X \odot \overline{\phi(X)})\right\|_2\right], \tag{16}$$

   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

$$H(\hat{F}_\phi) \approx \mathbb{E}\left[\|X \odot \phi(X)\|_2\right]. \tag{17}$$

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

$$H(\hat{F}_\phi) \approx \mathbb{E}\left[\|X \odot \phi(X)\|_1\right]. \tag{18}$$

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*. [71], 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 7)	**52.2**	**66.0**
GatedConv [71]	40.3	55.8
GatedConv, coarse outputs [71]	41.6	56.8
GatedConv, ℓ_1 only [71]	43.7	59.0
GatedConv+Fine tuning, ℓ_1 only [71]	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*. [71], 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*. [71] 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 [71] 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).