A number of problems in Computer Vision – image segmentation, geometric labeling, human body pose estimation – can be written as a mapping from an input image $x \in \mathcal{X}$ to an exponentially large space $\mathcal{Y}$ of structured outputs. For instance, in semantic segmentation, $\mathcal{Y}$ is the space of all possible (super-)pixel labelings, $|\mathcal{Y}| = L^n$, where $n$ is the number of (super-)pixels and $L$ is the number of object labels that each (super-)pixel can take.

As a number of empirical studies have found [4, 8, 13], the amount of training data is one of the most significant factors influencing the performance of a vision system. Unfortunately, unlike unstructured prediction problems – binary or multi-class classification – data annotation is a particularly expensive activity for structured prediction. For instance, in image segmentation annotations, we must label every (super-)pixel in every training image, which may easily run into millions. In pose estimation annotations, we must label every (super-)pixel in every training image, which may easily run into millions. In pose estimation annotations, we must label every (super-)pixel in every training image, which may easily run into millions.

As illustrated in Fig. 1, each bin in the histogram corresponds to a subset of solutions – for instance, all segmentations where size of foreground (number of ON pixels) is in a specific range $[L U]$. Computing the entropy of this coarse distribution is simple since $M$ is a small constant ($\approx 10$). Importantly, we prove that the optimal histogram, i.e. one that minimizes the KL-divergence to the Gibbs distribution, is composed of the mass of the Gibbs distribution in each bin, i.e. $\sum_{y \in \text{bin}} P(y|x)$. Unfortunately, the problem of estimating sums of the Gibbs distribution under general hamming-ball constraints continues to be #P-complete [11]. Thus, we upper bound the mass of the distribution in each bin with the maximum entry in a bin multiplied by the size of the bin. Fortunately, finding the most probable configuration in a hamming ball has been recently studied in the graphical models literature [1, 7, 9], and efficient algorithms have been developed, which we use in this work.

We perform experiments on figure-ground image segmentation and coarse 3D geometric labeling [5]. As shown in Fig. 2, our proposed approach significantly outperforms a large number of baselines and can help save hours of human annotation effort.

---

**References**


