Scalable Model-Assisted Multi-Target Estimation in Large Image Collections

Max Hamilton **Jinlin Lai** **Daniel Sheldon** **Subhransu Maji**

University of Massachusetts, Amherst

Abstract

Computer vision models are increasingly used as measurement tools to estimate population-level quantities from large image collections, but prediction errors introduce bias and the resulting estimates lack statistical guarantees required in scientific applications. Prior work uses a Monte Carlo framework to combine model predictions with ground-truth annotations by sampling some images for humans to label and is able to provide unbiased estimates with controllable accuracy, but primarily addresses single-scalar estimation. We study the more general problem of multi-target estimation, where many quantities (e.g., class counts or proportions) must be estimated simultaneously, and adapt sampling and estimation strategies from survey sampling to this setting. Evaluations on five detection and segmentation datasets with 7–80 classes show that importance sampling excels with moderate annotation budgets or fewer targets, whereas uniform sampling with control variates is superior when estimating many targets or operating with minimal labels. Additionally, a subset-based ratio estimator remains highly competitive across all regimes. Ultimately, our framework effectively combines biased model predictions and limited human labels into rigorous scientific measurements.

1 INTRODUCTION

Modern computer vision systems such as object detectors and image segmentation models are increasingly used as measurement tools over large image collections. These pipelines enable scientists to compute population-level quantities that would otherwise be prohibitively expensive to obtain manually. Examples include estimating species counts from aerial imagery [Belotti et al., 2023, Van Horn et al., 2018, Wu et al., 2023], measuring land-cover proportions from satellite data [Helber et al., 2019, Turkoglu et al., 2021, Robinson et al., 2019], and quantifying coral reef composition from underwater photographs [Sauder et al., 2025a,b].

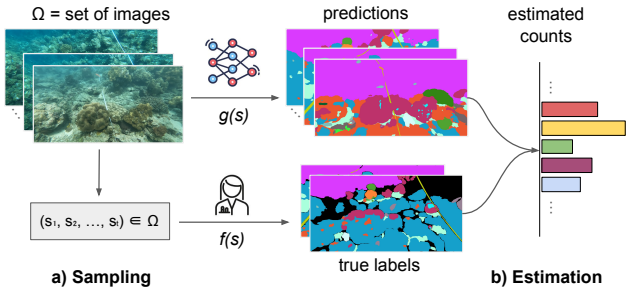

Figure 1: Our method combines model predictions $g(\cdot)$ with a small number of human annotations $f(\cdot)$ to obtain total counts of multiple targets in an image collection Ω. In the above example, a segmentation model predicts pixel labels for each class within each image in the Coralscapes dataset [Sauder et al., 2025b]. We explore *(a)* sampling strategies for human labeling and *(b)* statistical estimation techniques for obtaining total pixel or object counts for each class within the dataset.

Despite their practical utility, such measurements are often unreliable when models are deployed in challenging environments or novel domains. Prediction errors introduce unknown biases, and the resulting estimates lack statistical guarantees. In contrast, scientific applications typically require measurements with quantifiable uncertainty. To address this gap, recent work (e.g., [Meng et al., 2022, Perez et al., 2024, Angelopoulos et al., 2023a]) has proposed combining model predictions with a limited set of ground-truth annotations within a statistical estimation framework: the resulting estimates are unbiased, equipped with confidence intervals, and error goes to zero with the number of annotations. However, existing approaches largely focus on estimating a single scalar quantity—for example, the total

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:1931–1946.

count of one animal species across a large image collection.

In this paper we consider the more general and practically relevant problem of multi-target estimation, where the goal is to estimate a vector of quantities simultaneously (e.g., counts of many species, or class proportions; cf. Figure 1). Sampling strategies that work well for one target may perform poorly for others, and in general it is unknown how best to draw samples and construct estimators for multiple concurrent targets.

We study this problem by systematically adapting a range of sampling and estimation techniques from the survey sampling literature to our setting of computer vision measurement assisted by limited human annotation. We evaluate these estimators across five computer vision datasets spanning object detection and image segmentation tasks with 7–80 classes and 500–5000 images (Table 1). We also investigate how model accuracy and dataset composition impact the choice of estimator.

Our analysis reveals four key findings: (1) Importance sampling-based approaches (e.g., Perez et al. [2024]) excel when the number of targets is small and are effective at reducing worst-case per-class errors. (2) As the number of targets scales, uniform sampling augmented with control variates becomes highly robust and yields substantially lower error. (3) A subset-based ratio estimator performs competitively across all datasets and annotation budgets. (4) Across tasks, combining model predictions with limited ground-truth labels dramatically reduces measurement error, typically achieving less than 15% average fractional error across classes with 10% labeled data.

Overall, our results demonstrate that integrating model predictions with annotations within a statistical estimation framework enables substantially more accurate multi-target estimation than relying on human annotations alone, and that the optimal estimator depends on the accuracy of the model and the dimensionality of the measurement target.

Code for this paper is available at:

`https://github.com/cvl-umass/multi-measurement`

2 FORMULATION AND APPROACH

Figure 1 provides an overview of our proposed framework. To motivate the multi-target estimation problem, consider a marine biologist seeking to understand the distribution of coral species across a vast reef system. The biologist collects thousands of high-resolution underwater photographs and wishes to measure the population-level prevalence of each species. Obtaining exact counts by manually annotating every image is prohibitively expensive and time-consuming. On the other hand, deploying an off-the-shelf AI model to automatically predict species counts can introduce unknown biases, as the model may generalize poorly to new environmental conditions, particularly for rare or visually complex classes.

We formalize this setting as a multi-class measurement problem. Let Ω denote the image collection. For each unit $s \in \Omega$ and class $c \in \{1, \ldots, m\}$, there is an unknown ground truth measurement $f_c(s) \geq 0$. Labeling a given unit s yields a ground truth vector $f(s) = [f_1(s), \ldots, f_m(s)]^\top$. Our goal is to estimate the total measurement across the dataset, given by:

$$F(\Omega) = \sum_{s \in \Omega} f(s) \in \mathbb{R}^m.$$

Rather than labeling every image, we sequentially sample and label a small subset $s_1, s_2, \ldots, s_t$ and construct corresponding population estimates $\hat{F}_t$. To improve efficiency, we leverage predictions from an AI model. Let $g_c(s) > 0$ denote the predicted measurement for each unit $s \in \Omega$ and class $c \in \{1, \ldots, m\}$, and similarly $g(s) \in \mathbb{R}^m$. We will assume that the predicted measurements are known for all samples in advance. Denote the total measurement from the AI model as $G(\Omega) = \sum_{s \in \Omega} g(s) \in \mathbb{R}^m$. While $G(\Omega)$ can be computed cheaply, it may be biased. Our objective is to combine the exact but sparse human annotations $f(s)$ with the fast but potentially biased predictions $g(s)$ to obtain accurate and statistically grounded estimates of $F(\Omega)$.

The key design questions are: *(a)* how to select the sequence of samples $s_1, \ldots, s_t$, and *(b)* how to combine model predictions with true labels to obtain an estimate. Uniform sampling treats all units equally and provides a robust baseline that ignores model predictions. To improve label efficiency, non-uniform strategies such as importance sampling prioritize units proportional to their predicted measurements. However, multi-class estimation introduces a new challenge: an image that is highly informative for one class may be uninformative for another. To address this, we explore class-independent sampling strategies with control variates and hybrid schemes such as Midzuno–Sen sampling, which combine the targeted efficiency of importance sampling with the shared coverage of uniform sampling. We also explore various estimator choices, including the difference, ratio, and regression estimators, which can be used with different sampling schemes. These techniques are described in more detail in the next section, and we refer readers to a comprehensive review of foundational Monte Carlo methods in Owen [2013].

Overall, our framework unifies sampling design and estimator construction to enable accurate, multi-target measurement while labeling only a small fraction of the dataset.

2.1 UNIFORM SAMPLING

In order to form an estimate for a particular class, we need to label a new sample for that class. With a uniform sampling

scheme, all the classes have the same sampling distribution, so a sample for one class is valid for all. Thus we only need to label one sample which can then be used to compute estimates for all classes.

Monte Carlo The simplest unbiased baseline would be a simple Monte Carlo estimator,

$$\hat{F}_t^{\mathrm{MC}} = \frac{n}{t}\sum_{i=1}^{t} f(s_i)$$

where we multiply the sample mean by $n = |\Omega|$ to estimate the population sum. This estimate does not use any predicted counts. For this estimator we use simple random sampling without replacement (SRSWOR), with each unit being equally likely to be selected. Sampling without replacement avoids duplicate samples, reducing variance, and is particularly helpful when a large portion of units are labeled. In survey sampling, this estimator is equivalent to the Horvitz-Thompson estimator [Horvitz and Thompson, 1952] equipped with simple random sampling.

Difference We can leverage the predicted counts $g(s)$ from an AI model to reduce the variance even further. For example, one can use the difference estimator [Cochran, 1977, Chapter 7.1],

$$\hat{F}_t^{\mathrm{Diff}} = \frac{n}{t}\sum_{i=1}^{t} (f(s_i) - g(s_i)) + G(\Omega)$$

where $G(\Omega) = \sum_{s\in\Omega} g(s)$. The closer our predicted counts are to $f(s)$, the smaller the error. Importantly, this estimator is always unbiased, even if the AI model predictions are biased. Like with Monte Carlo, this estimator uses SRSWOR. The difference estimator can be interpreted as using model predictions as a control variate, a well-known variance-reduction technique, and also the basis of approaches like prediction-powered inference (PPI) [Angelopoulos et al., 2023a] (See § 3 and Owen [2013]).

Regression While the difference estimator corrects for a constant offset in the predicted counts, the regression estimator [Cochran, 1977, Chapter 7.2] generalizes this approach to correct for both shift and scaling errors. For any given class c, the estimator is defined as:

$$\hat{F}_{t,c}^{\mathrm{Reg}} = \frac{n}{t}\sum_{i=1}^{t}(f_c(s_i) - \beta_c^T g(s_i)) + \beta_c^T G(\Omega)$$

where β_c represents a per-class weight vector.

Rather than solely using the predicted counts for the specific class being estimated, $g_c(s_i)$, we incorporate the counts of all classes, denoted as $g(s_i)$. Consequently, our per-class weights form a vector $\beta_c \in \mathbb{R}^m$. This vector represents a linear combination of counts that approximates the ground truth $f_c(s_i)$.

To solve for the optimal β_c that minimizes overall variance, we perform least squares regression on the sampled units. This regression is recomputed every 10 iterations, yielding a continually updated, unique weight vector for each class. To simplify this calculation, we sample uniformly with replacement. Because β_c is derived from the same samples used to compute the estimate itself, $\hat{F}_{t,c}^{\mathrm{Reg}}$ does have a bias. However, this bias is asymptotically small [Cochran, 1977, Chapter 7.7]. A complete discussion detailing the computation of each β_c is provided in Appendix A.1.

2.2 IMPORTANCE SAMPLING

Importance sampling is a standard variance reduction technique in which units are sampled with a probability proportional to their predicted counts. By prioritizing regions of the data that contribute most heavily to the final estimate, this approach drastically reduces the variance without requiring a larger sample size. However, applying this technique to a multi-class setting introduces a challenge: an item that is highly 'important' for estimating one class might be irrelevant for another.

Round-Robin Importance Sampling A straightforward unbiased baseline to address this is round-robin importance sampling, where each class c is assigned its own dedicated subset of samples. Let $n_c = \lfloor t/m \rfloor + 1$ denote the number of samples allocated to class c after t total steps, and let $s_i^{(c)}$ denote the i-th sample drawn specifically for this class. The estimator is defined as:

$$\hat{F}_{t,c}^{\mathrm{RR}} = \frac{G_c(\Omega)}{n_c}\sum_{i=1}^{n_c}\frac{f_c(s_i^{(c)})}{g_c(s_i^{(c)})}$$

Here, each sample $s_i^{(c)}$ is drawn from a unique distribution proportional to g_c with replacement. While simple to implement, the round-robin scheme reduces the effective number of samples per class by a factor of m (the total number of classes). As a result, this approach scales poorly to tasks with high class counts.

Mixture Importance Sampling To improve sampling efficiency, we can instead draw (with replacement) from a single, shared mixture distribution, q, for all classes. This approach is sample efficient, allowing us to update estimates for every class using a single sample, and remains unbiased. Using t samples drawn from q, the estimator is given by:

$$\hat{F}_t^{\mathrm{Mix}} = \frac{1}{t}\sum_{i=1}^{t}\frac{f(s_i)}{q(s_i)}$$

While various choices exist for the shared mixture distribution q, we specifically select the distribution that would minimize the sum of variances across all classes if the predictions were perfect. This yields: $q(s) \propto \sqrt{\sum_{c=1}^{m} g_c(s)^2}$.

The full derivation of this result is provided in Appendix B. In practice, this formulation reduces the likelihood of drawing empty samples, providing more informative units for labeling.

Mixture Regression We can further reduce variance by adding regression. This yields the mixture regression estimator,

$$\hat{F}_{t,c}^{\text{MixReg}} = \frac{1}{t}\sum_{i=1}^{t}\frac{f_c(s_i) - \beta_c^T g(s_i)}{q(s_i)} + \beta_c^T G(\Omega)$$

Here, the samples s_i are drawn from the shared mixture distribution q, and the coefficients β_c are determined via per-class least squares regression (see Appendix A.1). This formulation combines the advantages of regression estimators and non-uniform sampling: it actively corrects for systematic shift and scale errors in the predictions while simultaneously avoiding the costly labeling of units with low expected counts. Following the same setup as the uniform regression estimator, it is important to note that this estimator introduces a small bias, though it vanishes asymptotically.

2.3 MIDZUNO-SEN SAMPLING

As we have discussed, relying on a single, shared sampling distribution for every class requires a compromise: the most important units for one class are often irrelevant for another. While round-robin sampling provides per-class sampling distributions, it suffers from poor sample efficiency.

To resolve this, we propose a method that has both per-class sampling distributions and high sample efficiency. We achieve this utilizing a specialized survey sampling technique: Midzuno-Sen (MS) sampling [Midzuno, 1951, Sen, 1953]. MS sampling operates as a two-phase hybrid of importance and uniform sampling: The first unit is drawn with a probability proportional to its predicted count, exactly as in importance sampling, and the remaining $t-1$ units are drawn uniformly without replacement from the rest of the dataset. Because only the first draw is specialized to a specific class, the remaining uniform draws can be shared to update the estimates for all other classes. This makes the approach sample efficient.

Crucially, MS sampling possesses a unique statistical property: the probability of selecting any specific subset of units (independent of the draw order) is proportional to the sum of the predicted counts within that subset. This property allows us to reframe our estimation as sampling over subsets rather than individual units, which directly motivates our choice of estimator.

Subset Estimator In standard survey sampling, the classical ratio estimator [Cochran, 1977, Chapter 6.2] is typically biased. However, Midzuno-Sen originally proposed this sampling scheme because its subset-probability property perfectly cancels out this bias, rendering the ratio estimator strictly unbiased. Therefore, we pair MS sampling with the ratio estimator for this multi-target task, which we refer to as the "Subset Estimator":

$$\hat{F}_t^{\text{Sub}} = G(\Omega)\frac{\sum_{i=1}^{t} f(s_i)}{\sum_{i=1}^{t} g(s_i)}$$

This gives the benefits of a good per-class proposal distribution while also being sample efficient. An implementation of this scheme can be found in Appendix D.

Subset Regression Estimator Finally, we can also add regression to this estimator,

$$\hat{F}_{t,c}^{\text{SubReg}} = G_c(\Omega)\frac{\sum_{i=1}^{t}(f_c(s_i) - \beta_c^T g(s_i))}{\sum_{i=1}^{t} g_c(s_i)} + \beta_c^T G(\Omega)$$

Solving for the optimal β_c would be computationally intractable, so we instead solve for the β_c that minimizes the variance of the first sample ($t = 1$). (see Appendix A.2).

3 RELATED WORK

There is a deep connection between our approaches and survey sampling: both require estimating a population quantity from a subset of units. In survey sampling, a common approach, called model-assisted estimation, is to use statistical models to guide the design of estimators [Särndal et al., 2003]. The estimators used in this work fall into this category. Under the framework of the generalized regression estimator [Cassel et al., 1976], the difference estimator, the regression estimator and the ratio estimator are derived using regression analysis of linear models. The subset estimator uses the Midzuno-Sen scheme [Midzuno, 1951, Sen, 1953], which makes the estimation with the ratio estimator unbiased. We adopt these approaches in scientific measurement problems by utilizing vision models for generating covariates in regression and building sampling distributions in non-uniform sampling.

Many machine learning methods share a similar idea of estimating the total with sampled data. In model evaluation, we often need to estimate a performance metric with a limited budget for labeling data. Active testing [Kossen et al., 2021] sequentially selects unlabeled data for human to label, according to a surrogate model of the performance metric, which builds an efficient estimator and updates the surrogate model at the same time. More related to this work, Fogliato et al. [2024] introduce simple survey sampling methods with uniform sampling, including the Horvitz-Thompson estimator [Horvitz and Thompson, 1952] and the difference estimator, to evaluate machine learning models under limited budget. In scientific measurement considered in this work, it is usually expensive to measure every unit with

Dataset	# Classes	# Samples	Task
Cityscapes	19	500	S
Coralscapes	39	552	S
COCO	80	5000	D
Chesapeake Landcover	7	1000	S
SACrop3K	10	1000	S

Table 1: A summary of the multi-class datasets considered in this work. For each dataset, we report the number of classes, the number of images, and whether it is a detection (D) or segmentation (S) task.

human efforts. In such a scenario, Perez et al. [2024] build an importance sampling estimator of the total measurement with a detector-based proposal distribution. Hamilton et al. [2025] combine detector-based importance sampling and active machine learning, achieving human-in-the-loop scientific measurement by estimating the total measurement while adapting the underlying predictive model. Compared with these works, our framework focuses on multiple estimands, which leads to different estimators and evaluations. The recently proposed prediction-powered inference (PPI) [Angelopoulos et al., 2023a,b] makes statistical estimates with labeled data and prediction on unlabeled data. Their estimator coincides with the difference estimator in survey sampling. Zrnic and Candès [2024] improve the sampling efficiency of PPI with non-uniform sampling, by designing sampling rules that consider uncertainty. Cowen-Breen et al. [2026] consider the budget allocation problem for PPI estimation with multiple predictions, while our focus is on multiple estimands. Our findings may suggest different ways of performing prediction-powered inference in scenarios with multiple estimands.

4 EXPERIMENTS

4.1 DATASETS AND MODELS

We perform experiments on five datasets comprising both image segmentation and object detection tasks (Table 1). These datasets span a range of class counts, dataset sizes, and levels of prediction accuracy. For segmentation datasets, we estimate the total pixel count for each class, which is useful for understanding relative abundance or land area in remote sensing applications. For detection datasets, we estimate the total number of objects for each class.

Cityscapes. As a standard segmentation benchmark, we employ the Cityscapes dataset [Cordts et al., 2016]. We use the test split, which consists of 500 images with fine pixel-level annotations captured in urban street scenes across 50 cities. We focus on the standard 19-class evaluation split. The predictions are obtained from SegFormer-B5 [Xie et al., 2021], a transformer-based segmentation model pretrained on the Cityscapes training set.

Coralscapes. To assess performance in a more specialized scientific domain, we utilize the Coralscapes dataset [Sauder et al., 2025b]. We use the combined validation and test splits, which contain a total of 552 high-resolution (1024×2048) images of coral reefs captured across 27 sites. The annotation structure is similar to that of Cityscapes but includes 39 distinct classes. The predictions are generated using a SegFormer-B2 model that has been fine-tuned on the Coralscapes training set.

COCO (Common Objects in Context). We include the Microsoft COCO dataset [Lin et al., 2015] to evaluate our method on a standard object detection benchmark. We use the val2017 split, which comprises 5,000 labeled images across 80 object categories. COCO features a highly diverse distribution of classes. The predicted counts are obtained from a Faster R-CNN model [Ren et al., 2016] with a ResNet-50-FPN backbone [He et al., 2016], trained on the COCO training set.

Chesapeake Landcover. To demonstrate applicability in high-resolution remote sensing, we utilize the Chesapeake Land Use and Land Cover dataset [Robinson et al., 2019]. This dataset provides 1-meter resolution classification maps derived from the USDA National Agriculture Imagery Program (NAIP) across the Chesapeake Bay watershed. It includes semantic categories such as water, forest, and impervious surfaces. For predictions, we use a U-Net model [Ronneberger et al., 2015] with a ResNet-50 encoder pretrained on ImageNet-1K v2 [Recht et al., 2019]. The model is fine-tuned on the training set for 50 epochs.

SACrop3K. Finally, we evaluate our framework on the South Africa Crop Type (SACrop3K) dataset [SACrop3K, 2025], a benchmark for agricultural monitoring using satellite time-series data. This dataset involves classifying field boundaries into 10 crop types (e.g., wheat, barley, and canola) using Sentinel-2 imagery. For predictions, we follow the same procedure as for Chesapeake, but fine-tune the model on the SACrop3K training set.

4.2 EVALUATION METRICS

We evaluate our estimators with mean fractional error across classes,

$$\frac{1}{m}\sum_{c=1}^{m}\frac{|\hat{F}_{t,c} - F_c(\Omega)|}{F_c(\Omega)}$$

where $\hat{F}_{t,c}$ is our estimate of class c with t samples and $F_c(\Omega)$ is the ground truth total count for class c.

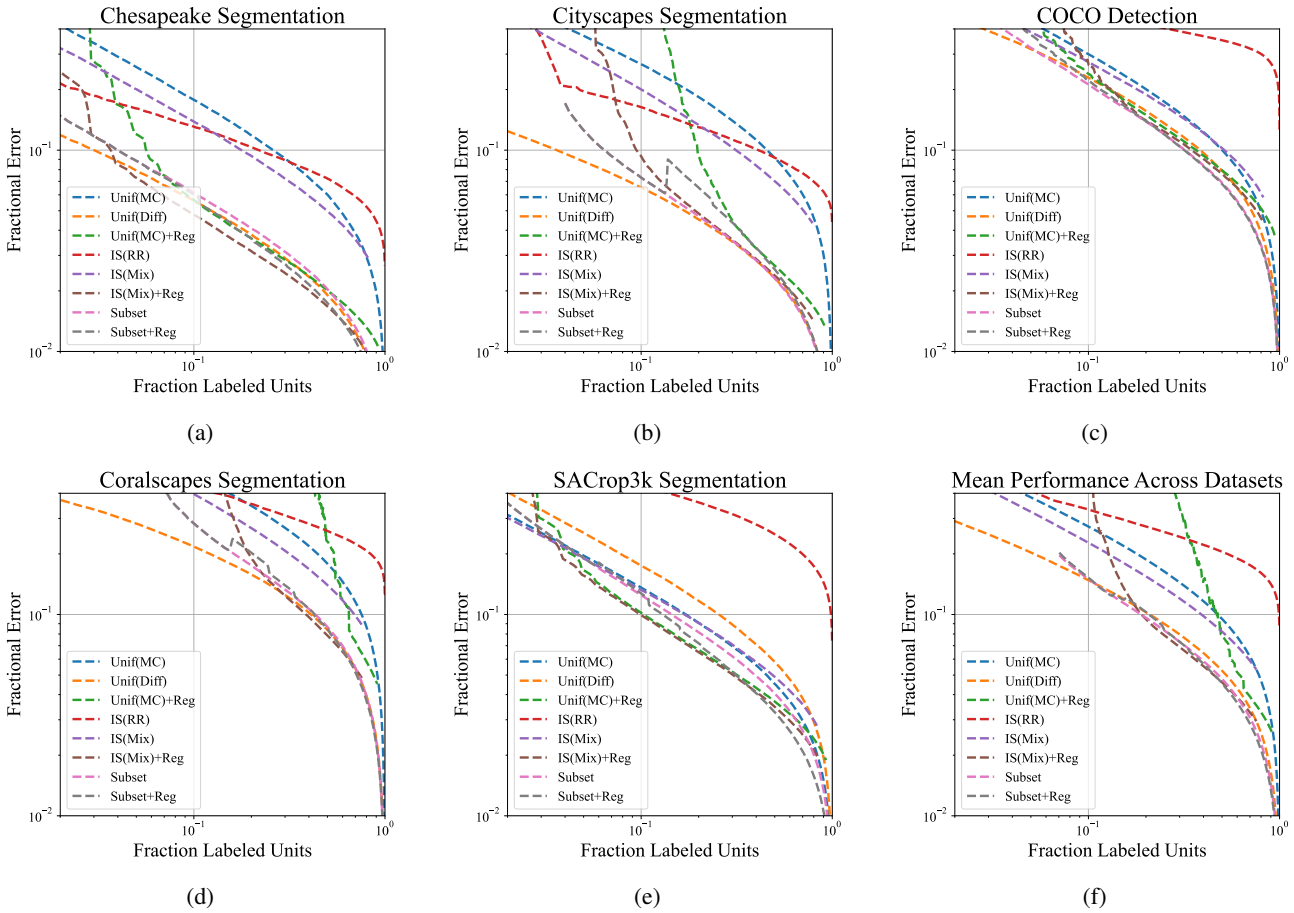

Figure 2: **Estimation error on 5 multi-target measurement tasks.** (a-e) Mean fractional error across all classes of the estimated per-class counts as a function of labeled units. (f) Performance of all methods averaged across the 5 datasets. We can see that Unif(Diff), IS(Mix)+Reg, and Subset estimators all perform well. Results are averaged over 5000 trials.

5 RESULTS

5.1 MAIN RESULTS

In Figure 2 we plot the mean fractional error as a function of labeled data for each dataset and method. Results are averaged over 5000 trials. For methods that sample with replacement, we plot against the number of unique labeled samples for a fairer comparison. We also report mean absolute error in Appendix E.

Overall, the Mixture IS + Regression, Subset, and Difference estimators are the best performing, though each excels under distinct conditions. We provide concrete recommendations for practitioners in Table 2. The Mixture Regression estimator typically performs best in the common scenario of 15%–50% labels. However, because it samples with replacement, its competitive edge diminishes at higher labeled percentages. It also lags at the start, as the regression requires a sufficient number of samples. Conversely, the difference estimators have the greatest advantage when fewer than 10% of labels are available and when the predictor is accurate. The SA Crop Type dataset is a good example of poor predictor performance, where the difference estimator underperforms the uniform MC baseline. Finally, the subset estimator delivers consistently solid performance across all settings and remains robust to errors in the predictions. Its main constraint is its reliance on per-class sampling distributions, requiring an initial number of samples equal to the total number of classes, m, before providing estimates.

Looking at the aggregate plot, we see that adding regression does not strictly improve the subset estimator. This is likely due to the added bias from the regression, as well as the fact that the regressed β values minimize variance at $t = 1$ rather than across all t, (a concession made to keep the regression objective tractable).

Comparing the uniform and non-uniform sampling, the mixture importance sampling consistently improves the estimate compared to uniform sampling. This holds true even in the difficult SA Crop Type dataset, where predictor performance is weaker. A similar trend can be seen with the regression estimators: the IS (Mix) + Reg estimator consistently con-

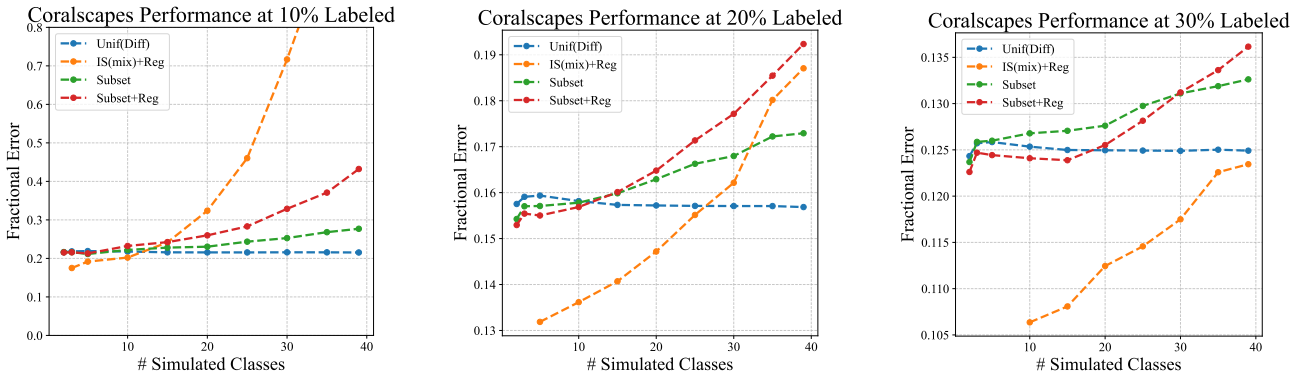

Figure 3: **Effect of class count on top estimators.** We simulate varying class counts on the Coralscapes dataset by averaging over uniform subsets of classes. The impact of class count diminishes as the number of labels grows.

verging to smaller error, and more quickly, than Unif (MC) + Reg. This is because uniform sampling often draws samples where both the predicted and ground-truth counts are 0, which fails to provide the regression with useful information, thereby requiring more total samples before converging to a good β.

Finally, the IS (RR) estimator has inconsistent performance relative to the Uniform estimator. This is largely due to the quality of the predicted counts. Round-robin sampling effectively reduces the number of samples for each class by a factor of m, which must be counteracted by a sufficiently good sampling distribution. If the predictions are not highly accurate, the variance is not reduced enough to counteract the factor of m fewer samples.

Table 2: Performance Summary of Key Estimators

Estimator	Regime	Strengths	Limitations
IS(Mix)+Reg	15%–50%	Converges to low mean error; reduces worst-case per-class error.	Lags initially (requires samples to fit β); more sensitive to large class counts.
Unif(Diff)	$< 15\%$	Highly efficient in low-label regimes; very robust to high class counts.	Relies on well-calibrated predictions; higher error spread.
Subset	All	Consistently reliable; robust to poor calibration and high class counts.	Requires initial sample size equal to class count (m); higher error spread.

5.2 EFFECT OF CLASS COUNT

We simulate the performance vs number of classes in Figure 3. Using the Coralscapes dataset, we average the error for each method over 5000 trials, with each trial getting a random subset of classes. The Difference estimator is very robust to the number of classes, showing little change in error. In contrast, the estimators using regression have the biggest increase in error as the number of classes grows, particularly Mixture Importance Sampling with Regression. This is likely because the mixture distribution becomes overly smooth with higher class counts, requiring more samples to converge to a good β. The subset estimator strikes a balance, being fairly robust in all cases. For all of these methods, the negative effect of increasing class size is greatly diminished as the percent of labeled data increases.

5.3 EFFECT OF PREDICTOR PERFORMANCE

We assess the impact of predictor quality in Figure 4. We run the estimators using predictions from three different checkpoints, representing models of different sizes trained on Cityscapes. Specifically, we use the B0, B2, and B5 configurations of SegFormer, containing 3.7M, 25.4M, and 82M parameters, respectively. Generally, error decreases as predictions improve with the larger models. The only exceptions are the Unif (MC) estimator, which does not incorporate predictions, and IS (Mix).

Looking closely at the Difference estimator, performance on the B0 and B2 models is significantly worse than on B5. The smaller B0 and B2 models struggle with visually complex classes, leading to substantial under-prediction. This further underscores the Difference estimator's reliance on accurate predictions. When model quality is uncertain or known to be poor, the Subset or IS(Mix)+Reg estimators should be preferred.

5.4 PERFORMANCE ACROSS CLASSES

Looking at the mean performance across classes may not convey the full story of an estimator's performance. To pro-

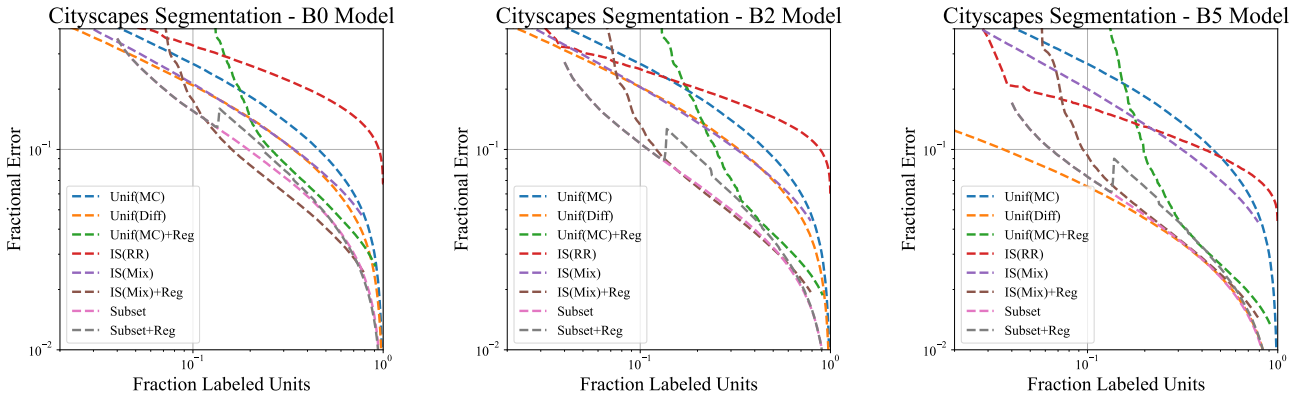

Figure 4: **Effect of prediction quality.** We estimate per-class counts on Cityscapes while varying the quality of predicted counts. The predicted counts come from SegFormer B0, B2, and B5, in order of increasing prediction quality and model size.

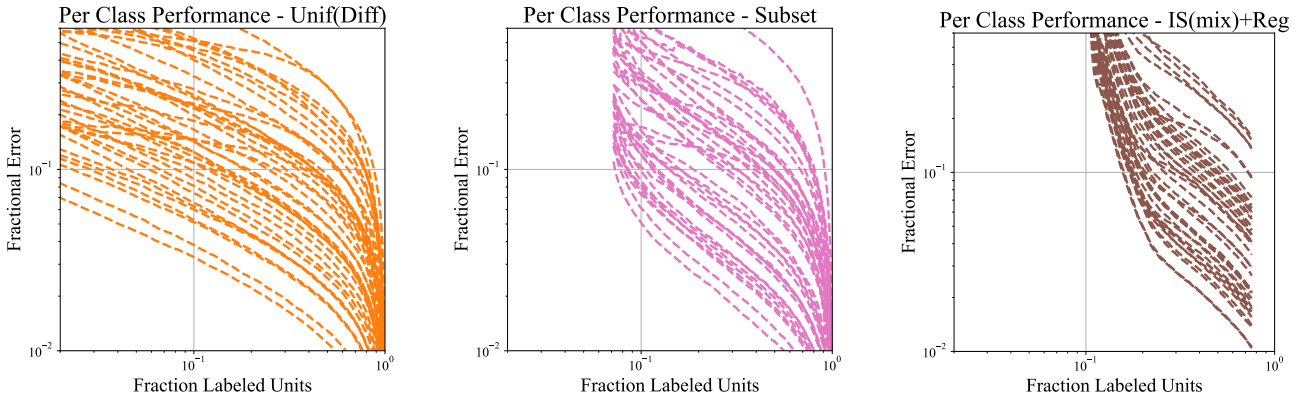

Figure 5: **Per-Class performance on Coralscapes.** Among top methods, IS(mix)+Reg has a tighter range of errors.

Table 3: Correlation between Coefficient of Variation and Fractional Error on Coralscapes Classes.

Method	**10% Labeled**	**20% Labeled**	**30% Labeled**
Unif(MC)	0.986	0.999	0.999
Unif(Diff)	0.562	0.580	0.584
Unif(MC)+Reg	0.155	0.301	0.402
IS(RR)	0.285	0.287	0.279
IS(Mix)	0.748	0.740	0.741
IS(Mix)+Reg	0.192	0.465	0.474
Subset	0.464	0.528	0.542
Subset+Reg	0.280	0.528	0.594

vide a more granular view, Figure 5 visualizes per-class errors for three representative estimators on the Coralscapes dataset. We provide visualizations for all classes in Appendix C.

These results reveal that the Difference and Subset estimators suffer from a significantly higher spread of errors across classes compared to the Mixture Regression estimator. Specifically, estimators utilizing uniform or subset sampling exhibit up to 30x more error in their worst-performing class compared to their best. In contrast, the Mixture Regression estimator limits this disparity to roughly 10x. For applications where minimizing the worst-case error is prioritized over optimizing the mean error, estimators utilizing Mixture Importance Sampling may be preferred.

To understand which factors determine the variation in per-class error, we analyze the correlation between the coefficient of variation (CV), which measures relative dispersion, and fractional error across all Coralscapes dataset classes. Looking at Table 3, we see that CV almost com-

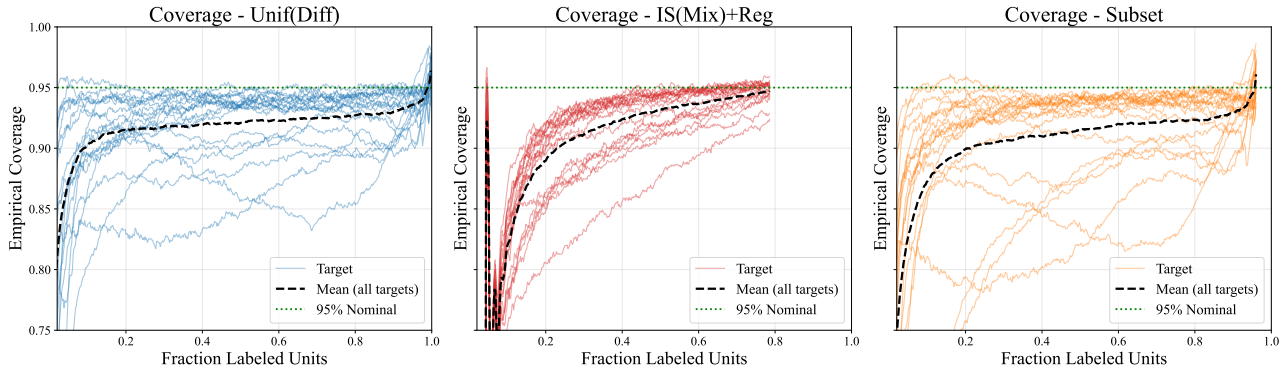

Figure 6: **Coverage of 95% confidence intervals on Cityscapes.** Per-class coverage for the best-performing estimators is robust across most classes, with under-coverage concentrated in sparse, high-variance classes.

pletely explains the variation in per-class fractional error for the Unif(MC) estimator ($r^2 \approx 0.99$), and explains approximately 25% of variation for methods that depend on predicted counts. The remaining 75% is likely driven by differing prediction qualities across classes, which play a substantial role in the per-class error of these methods.

5.5 CONFIDENCE INTERVALS

We construct variance estimates on a per-class basis using standard approaches from survey sampling [Owen, 2013], allowing us to form confidence intervals. Figure 6 displays the per-class coverage of 95% confidence intervals on the Cityscapes dataset for our best-performing estimators. Most classes achieve close to the target 95% coverage. However, we observe systematic under-coverage for rare classes with high variance, which is consistent with the per-class estimation error spread we observed earlier. IS(Mix)+Reg has better coverage in these worst-case scenarios.

6 CONCLUSIONS, LIMITATIONS, FUTURE WORK

This work demonstrates that integrating AI model predictions into multi-target estimation significantly enhances label efficiency. Through our evaluation of uniform and non-uniform sampling schemes, we found that the optimal estimator depends on operating constraints such as label budget, class count, and predictor calibration. The Difference estimator ($\hat{F}_t^{\text{Diff}}$) excels in highly constrained label regimes, provided the underlying model is well-calibrated. Conversely, the Subset estimator ($\hat{F}_t^{\text{Sub}}$) offers strong, consistent performance that is highly robust to calibration errors and scaling class counts. For standard labeling budgets between 15% and 50%, the Mixture Regression estimator ($\hat{F}_{t,c}^{\text{MixReg}}$) emerges as the strongest choice. Not only does it provide high overall accuracy, but it also reduces the spread of errors across classes. Ultimately, this work provides a practical framework to help researchers select the most effective estimation strategy by considering their specific data constraints and AI predictor capabilities.

A limitation of our work is that the estimates may not be accurate enough for all applications. This is especially true on a per-class basis, as our per-class results show significant variation in error rates across classes. One should be careful when deploying any of these methods in high-stakes scenarios. Furthermore, in this work we only explore per-class variance estimation, rather than the full covariance.

There are many promising directions for future work. First is the addition of model adaptation, where we finetune the predictor using the labels. This is easy to incorporate for the uniform sampling methods, as updating the model prediction only changes the computation of the estimate. On the other hand, changing the predictions for the Importance Sampling methods would additionally require updating the sampling distribution, adding a bit more complexity especially for Mixture Regression. Adaptation with the subset estimators can be done by only resampling the first sample proportional to the new predictions, as the uniform samples can be reused. This finetuning could also incorporate self-supervised or transductive learning.

Acknowledgements

We thank Rangel Daroya for providing technical support, access to her pre-trained model predictions, guidance on remote sensing datasets, and feedback on the draft. This work was supported in part by National Science Foundation grant #2504073.

References

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N Angelopoulos, John C Duchi, and Tijana Zrnic. PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

Maria Carolina TD Belotti, Yuting Deng, Wenlong Zhao, Victoria F Simons, Zezhou Cheng, Gustavo Perez, Elske Tielens, Subhransu Maji, Daniel Sheldon, Jeffrey F Kelly, et al. Long-term analysis of persistence and size of swallow and martin roosts in the US Great Lakes. *Remote Sensing in Ecology and Conservation*, 9(4):469–482, 2023.

Claes M Cassel, Carl E Särndal, and Jan H Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620, 1976.

William Gemmell Cochran. *Sampling techniques*. John Wiley & Sons, 1977.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Charlie Cowen-Breen, Alekh Agarwal, Stephen Bates, William W Cohen, Jacob Eisenstein, Amir Globerson, and Adam Fisch. Multiple-prediction-powered inference. In *The Fourteenth International Conference on Learning Representations*, 2026.

Riccardo Fogliato, Pratik Patil, Mathew Monfort, and Pietro Perona. A framework for efficient model evaluation through stratification, sampling, and estimation. In *European Conference on Computer Vision*, pages 140–158. Springer, 2024.

Max Hamilton, Jinlin Lai, Wenlong Zhao, Subhransu Maji, and Daniel Sheldon. Active measurement: Efficient estimation at scale. *Advances in Neural Information Processing Systems*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Jannik Kossen, Sebastian Farquhar, Yarin Gal, and Tom Rainforth. Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pages 5753–5763. PMLR, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context, 2015. URL `https://arxiv.org/abs/1405.0312`.

Chenlin Meng, Enci Liu, Willie Neiswanger, Jiaming Song, Marshall Burke, David Lobell, and Stefano Ermon. Iscount: Large-scale object counting from satellite images with covariate-based importance sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12034–12042, 2022.

Hiroshi Midzuno. On the sampling system with probability proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics*, 3(1):99–107, 1951.

Art B Owen. Monte Carlo theory, methods and examples, 2013.

Gustavo Perez, Subhransu Maji, and Daniel Sheldon. DISCount: counting in large image collections with detector-based importance sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22294–22302, 2024.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016.

Caleb Robinson, Le Hou, Kolya Malkin, Rachel Soobitsky, Jacob Czawlytko, Bistra Dilkina, and Nebojsa Jojic. Large scale high-resolution land cover mapping with multi-resolution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12726–12735, 2019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

SACrop3K. Planet, Radiant Earth Foundation, Western Cape Department of Agriculture, German Aerospace Center (DLR): A fusion dataset for crop type classification in Western Cape, South Africa. `https://source.`

coop/esa/fusion-competition, 2025. Accessed on 2025-03-03.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model assisted survey sampling*. Springer Science & Business Media, 2003.

Jonathan Sauder, Guilhem Banc-Prandi, Gabriela Perna, Ibrahim Souleiman Abdallah, Osama S Saad, Mustafa Altaib Mohammed, Ali Al-Sawalmih, Anders Meibom, and Devis Tuia. Rapid consistent reef surveys with deepreefmap. *Scientific Reports*, 15(1):39145, 2025a.

Jonathan Sauder, Viktor Domazetoski, Guilhem Banc-Prandi, Gabriela Perna, Anders Meibom, and Devis Tuia. The coralscapes dataset: Semantic scene understanding in coral reefs. In *Proceedings of the International Conference on Computer Vision Joint Workshop on Marine Vision*, 2025b.

Amode R Sen. On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5(1194):127, 1953.

Mehmet Ozgur Turkoglu, Stefano D'Aronco, Gregor Perich, Frank Liebisch, Constantin Streit, Konrad Schindler, and Jan Dirk Wegner. Crop mapping from image time series: Deep learning with multi-scale label hierarchies. *Remote Sensing of Environment*, 264:112603, 2021.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

Zijing Wu, Ce Zhang, Xiaowei Gu, Isla Duporge, Lacey F Hughey, Jared A Stabach, Andrew K Skidmore, J Grant C Hopcraft, Stephen J Lee, Peter M Atkinson, et al. Deep learning enables satellite-based monitoring of large populations of terrestrial mammals across heterogeneous landscape. *Nature communications*, 14(1):3072, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

Tijana Zrnic and Emmanuel J Candès. Active statistical inference. *arXiv preprint arXiv:2403.03208*, 2024.

Supplementary Material

Max Hamilton Jinlin Lai Daniel Sheldon Subhransu Maji

University of Massachusetts, Amherst

A REGRESSION PARAMETER DERIVATION

A.1 MIXTURE IS

Let $U(s) = f_c(s)/q(s)$ and $V(s) = g(s)/q(s) - G(\Omega)$. Since the samples are iid, the variance can be written as,

$$\text{Var}[\hat{F}_{t,c}^{\text{MixReg}}] = \frac{1}{t} E_{s\sim q}\left[\left(U(s) - \beta_c^T V(s) - F(\Omega)\right)^2\right]$$

Since each s_i is drawn from q, we can get an unbiased estimate using the batch of samples,

$$\approx \frac{1}{t}\sum_{i=1}^{t}(U(s_i) - \beta_c^T V(s_i) - F(\Omega))^2 \quad s_i \sim q$$

We can then solve for β_c and the offset $F(\Omega)$ using multiple linear regression to minimize the variance. The uniform regression estimator $\hat{F}_{t,c}^{\text{Reg}}$ is just a special case of this, where $q(s) = 1/n$.

A.2 SUBSET

For the subset estimator, which we can view as importance sampling over subsets, the optimal β_c would require solving a least squares problem over all t-subsets, which is computationally intractable. Additionally, we only have a single subset sample, so the batch estimate does not work. Instead, we solve for β_c which minimizes the variance of the first sample ($t = 1$). This makes it equivalent to the previous derivation except for the class-specific proposal distribution. Let $U(s) = G_c(\Omega)f_c(s)/g_c(s)$ and $V(s) = G_c(\Omega)g(s)/g_c(s) - G(\Omega)$. Then we have,

$$\text{Var}[\hat{F}_{1,c}^{\text{SubReg}}] = E_{s\sim q}\left[\left(U(s) - \beta_c^T V(s) - F(\Omega)\right)^2\right]$$

We can rewrite this in terms of an unweighted mean as,

$$\text{Var}[\hat{F}_{1,c}^{\text{SubReg}}] = \frac{1}{G_c(\Omega)} E\left[g_c(s_i)\left(U(s) - \beta_c^T V(s) - F(\Omega)\right)^2\right]$$

We can merge the g_c into the squared term by substituting, $Y(s) = G_c(\Omega)f_c(s)/\sqrt{g_c(s)}$ and $X(s) = G_c(\Omega)g(s)/\sqrt{g_c(s)} - \sqrt{g_c(s)}G(\Omega)$,

$$= \frac{1}{G_c(\Omega)} E\left[\left(Y(s) - \beta_c^T X(s) - \sqrt{g_c(s)}F(\Omega)\right)^2\right]$$

We can now get an unbiased estimate using the $t-1$ uniform samples from the Midzuno-Sen method,

$$\approx \frac{1}{G_c(\Omega)} \sum_{i=2}^{t}\left(Y(s_i) - \beta_c^T X(s_i) - \sqrt{g_c(s_i)}F(\Omega)\right)^2$$

If we add $\sqrt{g_c}$ as a column of X, this can then be minimized using multiple linear regression.

B DERIVATION OF MIXTURE IMPORTANCE SAMPLING

In mixture importance sampling, we select the distribution q that minimizes the sum of variances across all classes. Consider the importance sampling estimator $\hat{F} = f(s)/q(s)$ with a single sample s. The variance for class c is

$$\begin{aligned}\text{Var}[\hat{F}_c] &= \sum_{s\in\Omega} q(s)\left(\frac{f_c(s)}{q(s)} - F_c(\Omega)\right)^2 \\ &= \sum_{s\in\Omega} \frac{f_c(s)^2}{q(s)} - F_c(\Omega)^2.\end{aligned}$$

Then, the sum of variances is

$$\mathcal{V}(q) = \sum_{c=1}^{m} \text{Var}[\hat{F}_c] = \sum_{c=1}^{m}\sum_{s\in\Omega} \frac{f_c(s)^2}{q(s)} - F_c(\Omega)^2$$

subjected to the constraint $\sum_{s\in\Omega} q(s) = 1$. To minimize $\mathcal{V}(q)$, we apply the Lagrangian multiplier to the first part:

$$\mathcal{L}(q) = \sum_{c=1}^{m}\sum_{s\in\Omega} \frac{f_c(s)^2}{q(s)} + \lambda\left(\sum_{s\in\Omega} q(s) - 1\right).$$

For each $s \in \Omega$, $\frac{\partial \mathcal{L}(q)}{\partial q(s)} = 0$, which means

$$\begin{aligned} & -\sum_{c=1}^{m} \frac{f_c(s)^2}{q(s)^2} + \lambda = 0 \\ \iff \quad & q(s) = \sqrt{\frac{\sum_{c=1}^{m} f_c(s)^2}{\lambda}} \\ \iff \quad & q(s) \propto \sqrt{\sum_{c=1}^{m} f_c(s)^2}.\end{aligned}$$

Unfortunately, we do not have the ground-truth labels f beforehand. As a proxy, we substitute f with the predicted measurement g and set the proposal distribution as $q(s) \propto \sqrt{\sum_{c=1}^{m} g_c(s)^2}$.

C ADDITIONAL PER-CLASS RESULTS

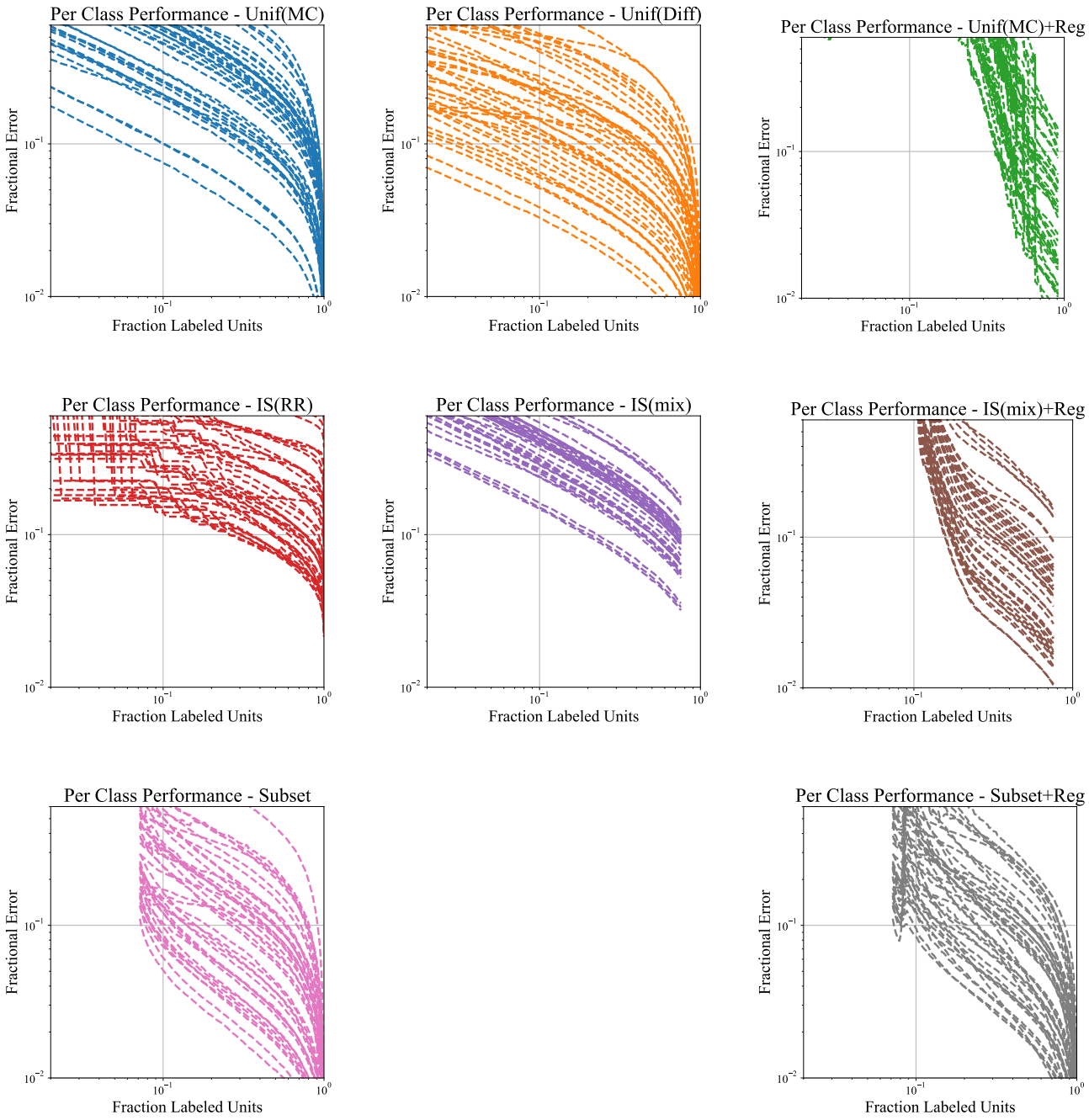

Figure 7: Per-Class performance of all methods on Coralscapes.

D IMPLEMENTATION OF SUBSET ESTIMATOR

Algorithm 1 Subset Estimation

Require: Population Ω, number of targets m. Predicted vectors $g(s) \in \mathbb{R}^m$ for all $s \in \Omega$. Labeling oracle returning true vectors $f(s) \in \mathbb{R}^m$.

Ensure: Sequence of estimated population sum vectors $\hat{F}_1, \hat{F}_2, \ldots$

▷ **Stage 1: Initial PPS Sampling**

$S_{\text{pps}} \leftarrow \emptyset$
for $c = 1$ **to** m **do**
  Sample s_c with probability $\propto g_c(s)$
  $S_{\text{pps}} \leftarrow S_{\text{pps}} \cup \{s_c\}$
  Query oracle to observe true labels $f(s_c)$
end for

▷ **Stage 2: Uniform Sampling & Estimation**

$S_{\text{uni}} \leftarrow \emptyset$
$t \leftarrow 1$
while stopping criterion is not met **do**
  Sample s_t uniformly without replacement from $\Omega \setminus (S_{\text{pps}} \cup S_{\text{uni}})$
  $S_{\text{uni}} \leftarrow S_{\text{uni}} \cup \{s_t\}$
  Query oracle to observe true labels $f(s_t)$
  $\hat{F}_t \leftarrow$ new vector in $\mathbb{R}^m$
  for $c = 1$ **to** m **do**
    $A_c \leftarrow S_{\text{pps}} \setminus \{s_c\}$ ▷ Stage 1 samples for other classes
    $T_{A,c} \leftarrow \sum_{s \in A_c} f_c(s)$ ▷ True sum of the "already labeled" set
    $G_{\text{rem},c} \leftarrow \sum_{s \in \Omega \setminus A_c} g_c(s)$ ▷ Predicted sum of remaining population
    $f_{\text{sum},c} \leftarrow f_c(s_c) + \sum_{s \in S_{\text{uni}}} f_c(s)$ ▷ Sampled true labels for remainder
    $g_{\text{sum},c} \leftarrow g_c(s_c) + \sum_{s \in S_{\text{uni}}} g_c(s)$ ▷ Sampled predictions for remainder
    $\hat{F}_{\text{rem},c} \leftarrow G_{\text{rem},c} \frac{f_{\text{sum},c}}{g_{\text{sum},c}}$ ▷ Ratio estimate for the remaining sum
    $\hat{F}_{t,c} \leftarrow T_{A,c} + \hat{F}_{\text{rem},c}$ ▷ Total population sum estimate for class c
  end for
  yield $\hat{F}_t = (\hat{F}_{t,1}, \ldots, \hat{F}_{t,m})$
  $t \leftarrow t + 1$
end while

E MEAN ABSOLUTE ERROR

Table 4: Mean Absolute Error on Coralscapes (pixels). Mean count: 24.78M pixels. IS(Mix)+Reg, Subset, and Unif(Diff) achieve the lowest MAE across most labeling percentages, consistent with fractional error rankings.

Method	**10% Labeled**	**20% Labeled**	**30% Labeled**
Unif(Diff)	1.84M	1.28M	1.00M
Subset	2.53M	1.38M	1.02M
IS(Mix)+Reg	50.01M	2.00M	1.25M
Unif(MC)	4.49M	3.05M	2.35M
IS(Mix)	5.74M	3.87M	3.00M
IS(RR)	5.54M	4.34M	3.65M
Subset+Reg	8.48M	2.83M	1.82M
Unif(MC)+Reg	578.26M	48.07M	21.02M