Conformal Prediction Sets for Instance Segmentation

Kerri Lu[1,2] **Dan M. Kluger**[4] **Stephen Bates**[1,2] **Sherrie Wang**[1,3,4]

[1]Laboratory for Information and Decision Systems, MIT, Cambridge, MA, USA
[2]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA
[3]Department of Mechanical Engineering, MIT, Cambridge, MA, USA
[4]Institute for Data, Systems, and Society, MIT, Cambridge, MA, USA

Abstract

Current instance segmentation models achieve high performance on average predictions, but lack principled uncertainty quantification: their outputs are not calibrated, and there is no guarantee that a predicted mask is close to the ground truth. To address this limitation, we introduce a conformal prediction algorithm to generate adaptive confidence sets for instance segmentation. Given an image and a pixel coordinate query, our algorithm generates a confidence set of instance predictions for that pixel, with a provable guarantee for the probability that at least one of the predictions has high Intersection-Over-Union (IoU) with the true object instance mask. We apply our algorithm to instance segmentation examples in agricultural field delineation, cell segmentation, and vehicle detection. Empirically, we find that our prediction sets vary in size based on query difficulty and attain the target coverage, outperforming baselines (naive best parameter and morphological dilation-based methods). We provide versions of the algorithm with asymptotic and finite sample guarantees. Our work is the first to capture structural uncertainty in instance segmentation by constructing confidence sets of diverse segmentation predictions.

1 INTRODUCTION

Instance segmentation aims to detect and delineate individual object instances in an image, producing a binary mask for each object. Despite achieving high performance on average predictions, current instance segmentation models lack principled uncertainty quantification: their outputs (such as logits or confidence scores) are not statistically calibrated, and there is no guarantee that a predicted mask is close to the ground truth. For example, the widely-used Segment Anything Model (Kirillov et al., 2023) outputs three binary masks for each query, along with their predicted Intersection-Over-Union (IoU) with the true mask, but there is no guarantee that any of these masks are correct or that their predicted IoUs are accurate. Similarly, ad-hoc uncertainty metrics, such as instance boundary probability predictions in agricultural field delineation (Waldner et al., 2021), are not calibrated and may be difficult to interpret.

Principled methods using conformal prediction have recently been created for segmentation, but remain constrained to modifications of a single model-predicted mask, e.g., through dilation or boundary expansion (Davenport, 2024; Mossina and Friedrich, 2025). These approaches capture local boundary uncertainty but cannot represent the structural ambiguity that dominates in many real applications, such as whether adjacent regions should be treated as a single object or split into multiple objects (see Appendix G for metrics quantifying this type of structural uncertainty). Addressing this gap requires confidence sets that contain qualitatively different masks, not just small perturbations of one mask.

To address this limitation, we introduce a conformal prediction algorithm to generate adaptive confidence sets for instance segmentation. Given an image and a pixel coordinate query, our algorithm generates a confidence set of instance predictions for that pixel, with a provable guarantee for the probability that at least one of the predictions has high IoU with the true object instance mask (Figure 1). Our algorithm (Section 2) generates a diverse set of predictions for each query by varying a tunable parameter of an instance segmentation model, and uses a randomly sampled calibration dataset to find a set of parameter values that is likely to result in high IoU predictions. Furthermore, we remove duplicate predictions to construct prediction sets that vary in size based on query difficulty. We provide versions of the algorithm with asymptotic and finite sample guarantees. We apply our algorithm to instance segmentation examples (Section 3) in agricultural field delineation, cell segmentation, and vehicle detection.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3999–4026.

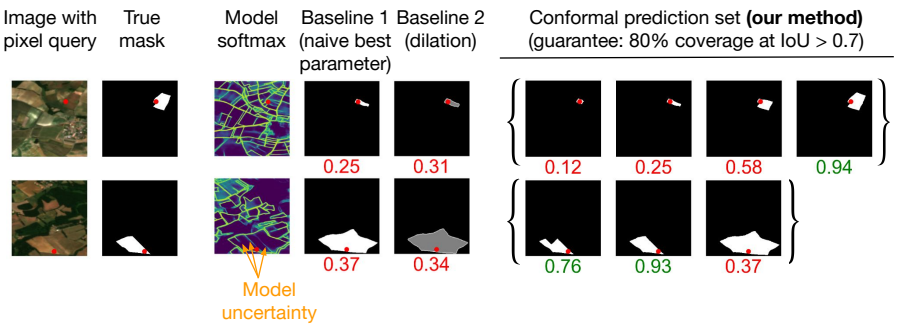

Figure 1: **Example field instance segmentation queries with true masks, model softmax scores, baseline predictions, and our method's conformal prediction sets (with IoU scores shown below each prediction).** Given an image and a pixel coordinate query, our conformal algorithm generates a confidence set of instance predictions for that pixel, with a provable guarantee for the probability that at least one of the predictions has high IoU (shown in green) with the true object instance mask. The naive best parameter baseline uses the single best model parameter value (over the calibration set) to generate a single prediction, but this often results in low IoU, as in the examples of over- and under-segmentation shown above. The dilation-based conformal baseline, which dilates the single prediction by a fixed number of pixels (determined using the calibration set), also fails to capture this structural ambiguity. By contrast, our method's confidence sets provide diverse predictions and adapt to query difficulty.

The main contribution of our work is a novel formulation of conformal prediction for instance segmentation that captures structural uncertainty. Specifically, (1) we identify structural uncertainty not captured by existing conformal approaches based on boundary perturbations; (2) we introduce a conformal prediction formulation that targets this setting by constructing set-valued predictions over multiple model configurations, enabling coverage guarantees even when no single parameter value can achieve the desired IoU threshold; and (3) we provide a practical instantiation by using a parameter sweep and set-cover-based selection, which allows us to probe structural diversity in existing models.

1.1 RELATED WORK

Conformal Prediction Conformal prediction methods generate confidence sets for machine learning model predictions, with provable coverage guarantees. Given an ML model and a labeled calibration dataset of randomly sampled points, conformal prediction uses a heuristic metric to measure uncertainty in model predictions and predict confidence sets for future unlabeled data points sampled from the same distribution (Angelopoulos, Bates, et al., 2023).

Two broad classes of conformal methods that have been applied to image segmentation are Conformal Risk Control (CRC) (Angelopoulos et al., 2022; Blot et al., 2024; Luo and Zhou, 2025; Luo et al., 2026) and Learn Then Test (LTT) (Angelopoulos et al., 2025b). Angelopoulos et al. (2022) introduced CRC and applied it to choose a model probability threshold for binary segmentation such that the expected false negative rate falls below a user-specified target. Angelopoulos et al. (2025b) introduced LTT and applied it to choose instance segmentation model parameters with guarantees for mean IoU, recall, and coverage.

Like our paper's method, both LTT and CRC are conformal methods that generate model predictions by varying a tunable parameter. However, these methods output a single prediction for each input, while our method outputs adaptive confidence sets with multiple predictions. The disadvantage of LTT and CRC is that there does not always exist a single parameter value that results in sufficiently high coverage on the calibration dataset. In our instance segmentation setting, there does not always exist a single parameter value that results in mask predictions with high IoU for a user-specified fraction $(1-\alpha)$ of calibration points. Thus, when attempting to guarantee high coverage on the test set, LTT or CRC can often return an error. In contrast, our method is feasible in more scenarios: it searches for a confidence set of multiple parameters such that with probability $(1-\alpha)$, predictions from *at least one* of the parameters will have high IoU.

CRC has the additional disadvantage of requiring a loss function that is monotone or near-monotone in the tunable parameter. Similarly, Risk-Controlling Prediction Sets (Bates et al., 2021), a precursor to LTT, also requires monotone loss. However, IoU loss is not monotone or near-monotone. In contrast, our method does not require a monotone loss function, which allows us to use an IoU-based loss function.

The general LTT framework described by Angelopoulos et al. (2025b) could potentially be adapted to our setting, but this requires greatly increasing the search space of tunable parameters. In our setting, if the tunable parameters are $t_1, ..., t_k$, then the LTT framework could be considered where the tunable parameters are elements of the power set $2^{\{t_1,...,t_k\}}$ (which has 2^k elements) rather than individual t_j values. However, this setting with an extremely large pa-

rameter space is not explored in the LTT paper, whereas we provide a prescriptive algorithm and implementation for producing a confidence set of multiple predictions.

We emphasize that our method extends the line of work introduced by the LTT and CRC frameworks: while prior instantiations of LTT/CRC select a single model configuration, we select a set of configurations whose predictions jointly achieve the desired coverage. Our method is the first to construct confidence sets of diverse segmentation predictions. Making the design choices that enable sets of predictions is non-trivial; we believe the construction of confidence sets rather than single predictions provides novelty and is necessary in order to capture structural uncertainty.

Uncertainty quantification for semantic and instance segmentation Several recent works use conformal methods to generate prediction sets for image segmentation tasks. For semantic segmentation, Mossina et al. (2024) used softmax scores to construct pixel-level confidence sets of class predictions, and Viti et al. (2025) developed image-level prediction sets that account for spatial correlations. These semantic segmentation methods do not distinguish between different object instances of the same class, and thus are not directly comparable to our method, which predicts a set of masks for each object instance.

For binary instance segmentation, Davenport (2024) combined predicted logit scores with boundary distances to construct inner and outer confidence sets, while Mossina and Friedrich (2025) applied morphological dilation to expand a predicted binary mask into a margin-shaped outer confidence set. However, these approaches only guarantee that the true mask is contained within the outer confidence set (and/or contains the inner confidence set) with high probability, and do not guarantee high IoU between the true mask and either confidence set. Furthermore, the confidence set is constrained to be a modification of an existing prediction, such as a dilation or boundary expansion. By contrast, our method constructs sets of alternative instance masks, where members of the set may differ substantially in shape. This allows us to capture more diverse and realistic forms of uncertainty. While our method's predictions are constrained to those generated by varying the tunable parameter, these are substantially more diverse (see Figure 2) compared to, e.g., dilating a single predicted mask by a fixed margin.

Uncertainty quantification for object detection Conformal methods have also been applied to object detection tasks, allowing users to control the false negative rate (Andéol et al., 2023, 2025; Li et al., 2022) or number of false positives (Fisch et al., 2022) in object bounding box predictions. Another approach is to use conformal prediction to construct confidence intervals for bounding box coordinates (De Grancey et al., 2022; Mukama et al., 2024; Timans et al., 2024; Zouzou et al., 2025). Bounding box predictions have limited flexibility and precision compared to mask predictions, which capture a wider range of object shapes.

2 METHOD

2.1 SETTING

We use conformal prediction to generate adaptive confidence sets for instance segmentation tasks. We consider the following setting:

- The input $X = (I, (z_1, z_2))$ consists of an image $I \in \mathbb{R}^{W \times H \times C}$ and a "query point" at pixel coordinate (z_1, z_2).
- Each query is located in a ground truth object instance, which can be denoted as a binary mask $Y \in \{0,1\}^{W \times H}$.
- We have access to f, an instance segmentation model that takes as input $X = (I, (z_1, z_2))$ and a tunable parameter T to output binary mask predictions $\hat{Y} = f(X, T)$. By varying T, we can modulate the behavior of f and obtain a diverse set of predictions. Examples of T include mask scores in Segment Anything, logit thresholds before connected-components analysis, and thresholds in watershed segmentation.
- To evaluate the quality of a mask prediction, we use the Intersection-over-Union metric,

$$\text{IoU}(Y, \hat{Y}) = \frac{|Y \cap \hat{Y}|}{|Y \cup \hat{Y}|},$$

  where a perfect prediction has IoU $= 1$. (For matrices $A, B \in \{0,1\}^{W \times H}$ we use $|A \cap B|$ and $|A \cup B|$ to denote the number of nonzero entries in the pointwise product $A \odot B$ and sum $A + B$, respectively).

This setting arises in applications ranging from farmers selecting their fields in satellite imagery to modern interactive models like Segment Anything, which take point prompts and return object masks. While we describe the method in the point-query setting, it extends to producing prediction sets for an entire image — either directly for models that support image-wide inference, or by querying all pixels in a brute-force manner. We adopt IoU as the evaluation metric for concreteness, though the method extends readily to other measures of segmentation quality.

Finally, to construct conformal prediction sets we require a calibration dataset $(X_1, Y_1), \ldots, (X_n, Y_n)$ consisting of n image-query pairs with their corresponding ground-truth masks. For a new test (or target) example $(X^{\text{test}}, Y^{\text{test}})$ coming from the same distribution, and given a user-specified IoU target $0 < \tau < 1$ and error rate $0 < \alpha < 1$, our conformal algorithm produces a confidence set $C_{\alpha,\tau}(X^{\text{test}})$. This confidence set is designed such that with probability $\geq 1-\alpha$,

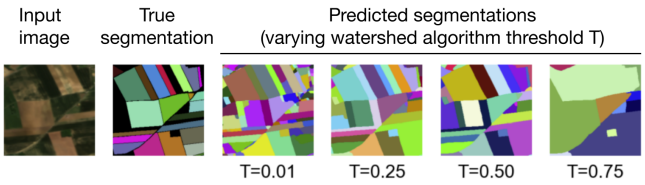

Figure 2: **Generating diverse predictions by varying tunable parameter T.** In the field delineation example, we vary the watershed algorithm threshold T to generate multiple segmentations for each image. While a given threshold may oversegment or undersegment a specific field, another threshold often succeeds, motivating the use of prediction sets that combine them.

it contains at least one predicted mask $\hat{y} \in C_{\alpha,\tau}(X^{\text{test}})$ such that $\text{IoU}(Y^{\text{test}}, \hat{y}) > \tau$. For formal theoretical guarantees, we assume the calibration and test samples are IID:

Assumption 1. $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} \mathbb{P}$, and independently, $(X^{\text{test}}, Y^{\text{test}}) \sim \mathbb{P}$ for all $n \in \mathbb{Z}_+$.

2.2 CONFORMAL INSTANCE SEGMENTATION ALGORITHM

Our conformal algorithm proceeds as follows (Algorithm 1). We begin by selecting a grid of k possible values for the tunable parameter T, denoted $\{t_1, t_2, \ldots, t_k\}$. Recall that T is any parameter that controls how f produces masks. Crucially, no single value of T yields the best prediction for all queries: at some settings, the model succeeds on certain queries, while other settings perform well on different queries (Figure 2). Our conformal algorithm leverages this variability by selecting a subset of T values that, together, provide coverage across the dataset.

For each calibration example (X_i, Y_i) and each value t_j (for $1 \le i \le n$ and $1 \le j \le k$), we use the instance segmentation model f to predict a binary mask $\hat{Y}_{ij} = f(X_i, t_j)$ and evaluate its quality using the IoU with the true mask: $\rho_{ij} = \text{IoU}(Y_i, \hat{Y}_{ij})$. Next, for each t_j, we collect the indices of calibration points where the prediction is sufficiently accurate:

$$S_j = \{i : \rho_{ij} > \tau\} = \big\{i : \text{IoU}\big(Y_i, f(X_i, t_j)\big) > \tau\big\}.$$

In other words, S_j is the set of calibration examples where $f(\cdot, t_j)$ achieves IoU above the threshold τ.

We then identify a minimal subset $J_{\alpha,\tau} \subseteq [k]$ of parameter values such that the union $\bigcup_{j \in J_{\alpha,\tau}} S_j$ covers at least $(1 - \alpha)n$ calibration examples. Intuitively, this ensures that the selected parameter values collectively succeed on a large fraction of the calibration set.

Finally, given a test image X^{test}, we generate predictions $\hat{Y}_j^{\text{test}} = f(X^{\text{test}}, t_j)$ for each $j \in J_{\alpha,\tau}$ and output the conformal prediction set

$$C_{\alpha,\tau}(X^{\text{test}}) = \{\hat{Y}_j^{\text{test}} : j \in J_{\alpha,\tau}\}.$$

Algorithm 1 Conformal prediction set for instance segmentation

Input:

- calibration data $(X_1, Y_1), \ldots, (X_n, Y_n)$
- test input X^{test}
- instance segmentation model $f(X, T)$
- parameter values $t_1, \ldots, t_k$
- error rate $0 < \alpha < 1$
- target IoU $0 < \tau < 1$

Output:

- parameter indices $J_{\alpha,\tau}$ for prediction set
- conformal prediction set $C_{\alpha,\tau}(X^{\text{test}}) = \{\hat{Y}_j^{\text{test}} : j \in J_{\alpha,\tau}\}$

for $j = 1, 2, \ldots, k$ **do**
  for $i = 1, 2, \ldots, n$ **do**
    $\hat{Y}_{ij} \leftarrow f(X_i, t_j)$
    $\rho_{ij} \leftarrow \text{IoU}(Y_i, \hat{Y}_{ij})$
  end for
  $S_j \leftarrow \{i : \rho_{ij} > \tau\}$
end for
$J_{\alpha,\tau} \leftarrow \arg\min_{J \subseteq [k]} \{|J| : \big|\bigcup_{j \in J} S_j\big| \ge (1 - \alpha)n\}$
$C_{\alpha,\tau}(X^{\text{test}}) \leftarrow \{f(X^{\text{test}}, t_j) : j \in J_{\alpha,\tau}\}$
return $C_{\alpha,\tau}(X^{\text{test}}), J_{\alpha,\tau}$

Computing a minimal-size set cover The step of identifying $J_{\alpha,\tau}$ is a variant of the classical Set Cover problem, which is NP-hard (Karp, 1972). To make this practical, we use a two-stage approach. First, we apply the greedy set cover algorithm (Chvatal, 1979; David S Johnson, 1973; Lovász, 1975), which runs in polynomial time, to obtain an initial cover $J'_{\alpha,\tau} \subseteq [k]$ that covers at least $(1 - \alpha)n$ calibration examples. We then attempt to improve this solution by brute-forcing over all subsets of $[k]$ of smaller size, ordered by cardinality, until we either find a true minimal cover or confirm that $J'_{\alpha,\tau}$ is already minimal. In practice, this refinement step is feasible because the greedy solution typically yields a small set. (However, if the greedy set cover $J'_{\alpha,\tau}$ is too large for the brute-force refinement step to be computationally tractable, we omit the refinement step.)

Conformal guarantee Because the test point is drawn from the same distribution as the calibration data, our procedure inherits a conformal coverage guarantee. In particular, with probability at least $(1 - \alpha)$ asymptotically, there is a predicted mask $\hat{y}$ in the confidence set $C_{\alpha,\tau}(X^{\text{test}})$ whose IoU with the true mask Y^{test} is greater than τ (see Proposition C.1). These guarantees rely on a feasible choice of α, τ.

We note that our method differs from standard conformal prediction in that Algorithm 1 does not use a non-conformity

score and the theory does not use exchangeability arguments. However, we still consider our method to fall under the conformal prediction framework because it uses a randomly sampled calibration set to generate statistically valid confidence sets for individual model predictions.

Not every α, τ are possible The quality of the conformal guarantees depends fundamentally on the underlying segmentation model f. If more than αn calibration points have no parameter $t_j \in \{t_1, \cdots, t_k\}$ that yields IoU greater than τ, then no subset $J_{\alpha,\tau}$ can cover the required $(1-\alpha)n$ calibration examples. In such cases, Algorithm 1 Line 8 fails to run. To address such issues, a user can either increase the error rate α, lower the IoU target τ, or consider a broader collection of segmentation models by considering a superset of the segmentation model parameters $\{t_1, \ldots, t_k\}$.

2.3 ADAPTIVE PREDICTION SETS VIA DUPLICATE REMOVAL

For any test input X^{test}, Algorithm 1 outputs a confidence set $C_{\alpha,\tau}(X^{\text{test}})$ whose size is fixed at $|J_{\alpha,\tau}|$. In practice, however, many of the masks $\hat{Y}_j^{\text{test}}$ for $j \in J_{\alpha,\tau}$ may be identical or highly similar, depending on the underlying model. To avoid redundancy, we post-process $C_{\alpha,\tau}(X^{\text{test}})$ using Algorithm 2 to remove near-duplicate masks. This yields confidence sets whose sizes adapt to the particular input X^{test}, providing a more informative representation of uncertainty.

Adaptive prediction set Formally, let $0 < \eta < 1$ be a user-specified IoU threshold at which two masks are considered duplicates. We define

$$C_{\alpha,\tau,\eta}(X^{\text{test}}) \subseteq C_{\alpha,\tau}(X^{\text{test}})$$

as a minimal-size subset such that for each mask prediction $\hat{Y} \in C_{\alpha,\tau}(X^{\text{test}}) \setminus C_{\alpha,\tau,\eta}(X^{\text{test}})$, there exists $\hat{Y}' \in C_{\alpha,\tau,\eta}(X^{\text{test}})$ satisfying $\text{IoU}(\hat{Y}, \hat{Y}') > \eta$. In other words, the new set is a minimal subset such that every discarded mask is within IoU η of one that is kept.

Computing this minimal subset is equivalent to the Dominating Set problem in graph theory, which is also NP-hard (Garey and David S. Johnson, 1979). In our implementation, we iterate over all possible subsets $C_{\alpha,\tau,\eta}(X^{\text{test}}) \subseteq C_{\alpha,\tau}(X^{\text{test}})$ in order of increasing cardinality until we find one that satisfies the above condition. This brute-force search is tractable in our experiments because $|C_{\alpha,\tau}(X^{\text{test}})| = |J_{\alpha,\tau}|$ is small. For larger set sizes, exhaustive search would quickly become infeasible, and one would need to resort to approximation or heuristic algorithms developed for the Dominating Set problem.

New conformal guarantee After removing duplicates, the new prediction set $C_{\alpha,\tau,\eta}(X^{\text{test}})$ will not necessarily satisfy the original conformal guarantee. To recover a valid guarantee, we re-calibrate the IoU threshold. For each calibration point (X_i, Y_i), we construct the adaptive set $C_{\alpha,\tau,\eta}(X_i)$ and record the best IoU achieved within that set:

$$s_i = \max_{\hat{Y}_i \in C_{\alpha,\tau,\eta}(X_i)} \text{IoU}(Y_i, \hat{Y}_i).$$

We then set $\tilde{\theta}$ to be the α-quantile of the scores $\{s_i : 1 \leq i \leq n\}$ across the calibration dataset. The resulting guarantee is that, for a new test input X^{test}, with probability at least $1-\alpha$, there exists some $\hat{Y}^{\text{test}} \in C_{\alpha,\tau,\eta}(X^{\text{test}})$ with $\text{IoU}(Y^{\text{test}}, \hat{Y}^{\text{test}}) > \tilde{\theta}$. This is stated formally in the following theorem, which is proven in Section C.4.

Theorem 2.1. *Let $C^{(n)}_{\alpha,\tau,\eta}(X^{\text{test}})$ and $\tilde{\theta}^{(n)}$ denote the set of segmentation masks and* IoU *threshold returned when running Algorithm 2 with calibration samples $\big((X_i, Y_i)\big)_{i=1}^n$ and parameters $\alpha, \tau, \eta \in (0,1)$. Under Assumptions 1, A.3, and A.4,*

$$\liminf_{n\to\infty} \mathbb{P}\Big(\max_{\hat{y} \in C^{(n)}_{\alpha,\tau,\eta}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}}, \hat{y})\big\} \geq \tilde{\theta}^{(n)} \Big) \geq 1-\alpha.$$

Theorem 2.1 holds under the IID assumption (Assumption 1), and two technical assumptions introduced in Section C.1.

Algorithm 2 Compute conformal prediction set for instance segmentation, with duplicates removed

Input:

- same inputs as Algorithm 1
- duplicate prediction IoU threshold $0 < \eta < 1$

Output:

- conformal prediction set with duplicates removed $C_{\alpha,\tau,\eta}(X^{\text{test}})$
- new IoU threshold $\tilde{\theta}$ for conformal guarantee

procedure UNIQUE(C, η)
  for $C' \subseteq C$ in increasing cardinality **do**
    if for all $\hat{Y} \in C \setminus C'$ there exists $\hat{Y}' \in C'$ such that $\text{IoU}(\hat{Y}, \hat{Y}') > \eta$ **then**
      return C'
    end if
  end for
end procedure

// remove duplicates from conformal prediction set
$C_{\alpha,\tau}(X^{\text{test}}), J_{\alpha,\tau} \leftarrow$ ALGORITHM-1
$C_{\alpha,\tau,\eta}(X^{\text{test}}) \leftarrow$ UNIQUE$(C_{\alpha,\tau}(X^{\text{test}}), \eta)$

// re-calibrate conformal guarantee IoU threshold
for $i = 1, 2, \ldots, n$ **do**
  $C_{\alpha,\tau}(X_i) \leftarrow \{f(X_i, t_j) : j \in J_{\alpha,\tau}\}$
  $C_{\alpha,\tau,\eta}(X_i) \leftarrow$ UNIQUE$(C_{\alpha,\tau}(X_i), \eta)$
  $s_i \leftarrow \max_{\hat{Y}_i \in C_{\alpha,\tau,\eta}(X_i)} \text{IoU}(Y_i, \hat{Y}_i)$
end for
$\tilde{\theta} \leftarrow \text{Quantile}_\alpha\{s_i\}_{i=1}^n$
return $C_{\alpha,\tau,\eta}(X^{\text{test}}), \tilde{\theta}$

These additional assumptions will typically hold, provided that the user makes appropriate implementation and input parameter choices.

While $\tilde{\theta}$ may differ from τ, we do not expect it to differ substantially as long as the duplicate removal IoU threshold η is high, since confidence sets that cover the true mask will often still cover the true mask after removing redundant predictions.

Finite sample guarantees The above theorem provides asymptotic guarantees. In Appendix H, we present a modified version of our algorithm with finite sample guarantees. We randomly split the calibration data into two sets; we use the first set to rank parameter values $t_1, t_2, \ldots t_k$ from highest to lowest priority via a greedy algorithm, then apply Conformal Risk Control to the second set to compute the smallest value $\hat{\lambda} \in [k]$ such that the first $\hat{\lambda}$ parameters in the ranked list result in a valid conformal prediction set. (The sample splitting allows for finite sample guarantees; in contrast, reusing the calibration data between Algorithms 1 and 2 breaks exchangeability and is a major reason we are only able to obtain asymptotic guarantees for Algorithm 2.)

The finite-sample and asymptotic algorithms will generally give similar results because the greedy procedure used in the finite-sample algorithm is closely related to the greedy set cover algorithm used in Line 8 of Algorithm 1. However, the asymptotic version's prediction sets may be smaller in some cases because in the asymptotic version (1) we can improve the greedy set cover by brute-forcing over all smaller subsets to find the minimal-size set cover, and (2) the duplicate removal procedure brute-forces over all subsets to produce a minimal-size output. In the finite-sample algorithm, because CRC requires a loss function that is non-decreasing in λ, the range of possible prediction sets is more restricted and we are unable to guarantee minimal-size outputs. We note that the finite-sample algorithm has the advantage of computational efficiency because its greedy parameter ranking procedure and duplicate removal procedure run in polynomial time, making it useful in cases where brute-forcing is not computationally tractable.

In our experiments in the following section, we use the asymptotic version of the algorithm. (For experimental results using the finite-sample version of the algorithm, see Appendix J.)

3 EXPERIMENTS

We evaluate our algorithm on three major use cases of instance segmentation where uncertainty quantification is important: agricultural field delineation, cell segmentation, and vehicle detection. For each dataset, the inputs consist of images paired with pixel-wise instance segmentation labels. All images were held out from the training of the segmentation models, i.e., they come from the models' test sets. We further divide these images into two disjoint subsets: a calibration set used to calibrate our conformal procedure, and a test set used to evaluate coverage and accuracy. For each image, we randomly sample pixel coordinates to serve as queries, and vary the tunable parameter T of the segmentation model to generate multiple instance predictions per query. We summarize the datasets, models, tunable parameters, and number of calibration and test pixels in Table 1. For detailed descriptions, see Appendix I.

Table 1: **Datasets and models used in our experiments.**

EXAMPLE	DATASET	MODEL	TUNABLE PARAMETER(S)	CALIB. PIXELS	TEST PIXELS
Field delineation	Fields of The World (Kerner et al., 2025)	FoTW UNet, watershed algorithm	watershed threshold $0 \leq T \leq 1$	2476	917
Cell segmentation	Cellpose (Stringer et al., 2021)	Cellpose-SAM (Pachitariu et al., 2025)	cell extent threshold $-5 \leq T \leq 5$	925	659
Vehicle detection	Cityscapes (Cordts et al., 2016)	Segment Anything Model (SAM) (Kirillov et al., 2023)	mask index $T_1 \in \{1, 2, 3\}$, probability threshold $0 \leq T_2 \leq 1$	105	121

Our procedure follows the framework described in Section 2. In each experiment, we calibrate our algorithm using the calibration queries, then construct $(1 - \alpha)$ confidence sets of instance predictions for each test query. To evaluate our outputs, we visualize the prediction sets, check that they satisfy the $(1 - \alpha)$ conformal coverage guarantee, and compare our results with baseline model predictions.

The choice of error rate α and IoU threshold τ is task-specific, reflecting the quality of the underlying segmentation model; the values for each experiment are reported in Section 3.2. For duplicate removal, we use $\eta = 0.9$ in all experiments; in practice, an end user would specify η based on what level of overlap would be considered a duplicate in their application. (See Appendix K for sensitivity analyses in which we vary α, τ, η, tunable parameter space size k, and calibration set size n.)

3.1 BASELINES

Naive best parameter baseline We create a "naive best parameter baseline" that always outputs the single best tunable parameter value, i.e. the parameter value that results in the highest fraction of predictions with IoU $> \tau$ over the calibration set. This baseline represents the best possible coverage we can obtain using a single parameter value. For the field delineation, cell segmentation, and vehicle detection calibration datasets, the single best parameter values are respectively $T = 0.24$, $T = 0$, and $T = (1, 0.05)$.

Morphological dilation baseline We also compare our method to the morphological dilation-based method from Mossina and Friedrich (2025). This method constructs con-

formal confidence sets for binary segmentation using dilation, i.e., adding a margin to a predicted mask, where the margin size is a fixed number determined using calibration data. This algorithm only guarantees that the true mask is contained within the dilated mask with high probability, and does not guarantee high IoU. We find that the dilation-based method results in low coverage for IoU, undersegmentation, and lack of flexibility in predictions compared to our method. See Appendix F for details on the experimental setup and results.

3.2 RESULTS

What α, τ are possible? For each experiment, we minimized the target error rate α while maximizing IoU threshold τ, subject to feasibility. To guide our choices, we plotted histograms of the maximum IoU achievable under any value of the tunable parameter T for each experiment (Figure D.1). This analysis revealed the inherent performance ceilings of the base models. For example, in the field delineation task, it was not possible to construct a 90% confidence set for IoU > 0.8, because even when taking the best T for each query, $> 10\%$ of predictions had IoU < 0.8. (We note that choosing α and τ by inspecting calibration histograms is a form of double-dipping in the calibration data, albeit a mild one as we do not expect it to have a large effect on the empirical coverage in practice. To avoid this, a user could prespecify α and τ; however, if the preselected α is too small for the preselected IoU threshold τ, the algorithm would return an error message as it is unable to guarantee $(1-\alpha)$ coverage.)

This feasibility analysis defines a Pareto frontier of achievable (α, τ) pairs: decreasing α (tighter error guarantees) necessarily requires lowering τ, and conversely, raising τ requires allowing larger α. We use $(\alpha = 0.2, \tau = 0.7)$ for fields, $(\alpha = 0.2, \tau = 0.75)$ for cells, and $(\alpha = 0.1, \tau = 0.8)$ for vehicles, choices that lie on the frontier and reflect the strictest feasible guarantees under the respective models (Table 2). Field delineation has the weakest guarantee, as it is the most ambiguous task among the three, while SAM vehicle segmentation has the strongest guarantee. Quantifying these ceilings shows how far current models are from reliability levels needed in practice, and it clarifies that conformal prediction can certify but not overcome the limits of the underlying model.

Conformal prediction set coverage Given these feasible (α, τ) pairs, our conformal algorithm achieves empirical coverage close to the target $1-\alpha$ and substantially higher than naive best parameter baselines (Table 2). Figure 3 shows cumulative distributions of IoUs for naive best parameter baseline predictions versus conformal sets, where for each test query we take the maximum IoU among the masks in the set. In field delineation, our conformal sets increase coverage from 66.3% to 79.7%, recovering many cases where the baseline either merged multiple fields or split one field incorrectly. In vehicle detection, coverage improves from 66.1% to 84.3%, demonstrating that alternative hypotheses often capture cars missed by the top SAM prediction. By contrast, cell segmentation shows only a modest improvement (80.3% to 83.5%), reflecting that most errors in this domain are boundary-local. When visualizing example conformal prediction sets, we see that the sets usually contain at least one mask that is very similar to the true mask (Figure 5). In general, using a confidence set of multiple parameters makes it more likely that at least one of the resulting predictions will have high IoU, compared to outputting a single prediction as the naive best parameter baseline does (see Appendix E for an illustrative example).

In all three examples, $\tilde{\theta} > \tau$ or $\tilde{\theta} \approx \tau$, so removing duplicate predictions does not significantly affect the original conformal guarantee. (Note that $\tilde{\theta} > \tau$ occurs when $\bigcup_{j \in J_{\alpha,\tau}} S_j$ contains more than $(1-\alpha)n$ calibration indices, so using the original IoU threshold τ would result in over-coverage.)

Table 2: **Target, Conformal, and Naive Best Parameter Baseline Coverages of Test Set Predictions.**

EXAMPLE	τ	$\tilde{\theta}$	TARGET COVERAGE $(1-\alpha)$	CONFORMAL COVERAGE	BASELINE COVERAGE (NAIVE BEST PARAMETER)
Field delineation	0.7	0.696	0.8	0.797	0.663
Cell segmentation	0.75	0.760	0.8	0.835	0.803
Vehicle detection	0.8	0.842	0.9	0.843	0.661

Adaptive set size In addition to achieving calibrated coverage, the experiments illustrate that our method adapts the size of prediction sets to query difficulty. Figure 4 shows the distribution of set sizes after duplicate removal: in most cases, sets contain three or fewer masks, but they expand when the query is ambiguous. This adaptivity improves efficiency, ensuring that users are not overwhelmed by redundant masks while still being presented with multiple plausible options in challenging cases. Field delineation has the largest set sizes, going up to 5 potential masks. The ambiguous field boundaries yield masks of significantly different sizes, corresponding to under- and over-segmentation hypotheses (Figure 5). In contrast, for cells, most sets collapse to size one or two, again consistent with the observation that errors are primarily boundary-local. (A few sets have size zero because the model classifies the queried pixel as a non-cell.) In vehicle detection, our method often recovers useful predictions beyond the top-1 mask returned by SAM. Together, these results show that conformal sets provide not only calibrated coverage but also an adaptive representation of uncertainty that reflects the difficulty of each query. This adaptivity and diversity of predictions is not available in the naive best parameter baseline, which always outputs a

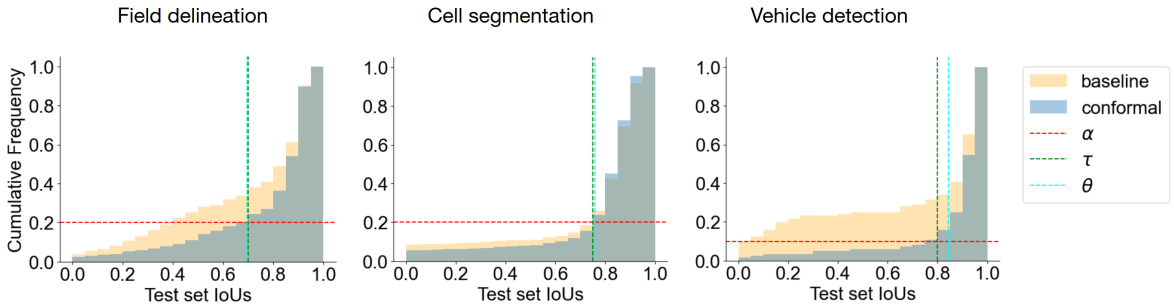

Figure 3: **Cumulative distributions of IoUs of naive best parameter baseline predictions and our conformal prediction sets over test data points.** For our conformal prediction sets, the maximum IoU over all predictions is used. In all examples, our conformal coverage for IoU $> \tilde{\theta}$ is close to target coverage $(1 - \alpha)$ and greater than baseline coverage. Furthermore, the re-calibrated IoU threshold $\tilde{\theta}$ is close to (or greater than) the original threshold τ, so removing duplicates does not significantly affect the original conformal guarantee. Note that the value of the baseline cumulative frequency at τ corresponds to the smallest error rate at which the naive best parameter baseline would provide coverage guarantees.

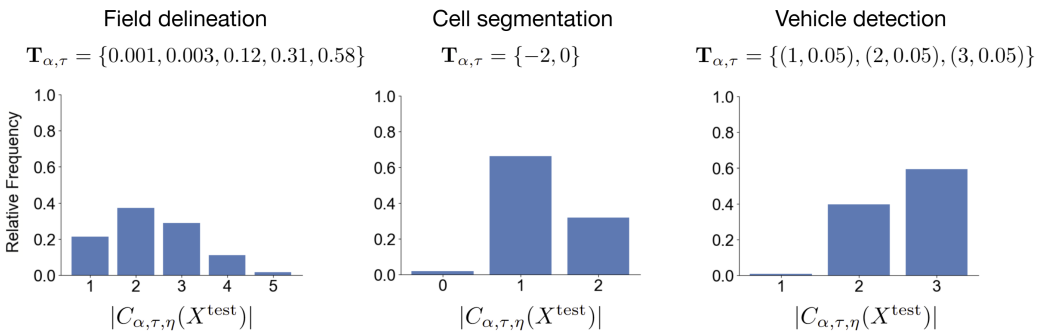

Figure 4: **Distribution of conformal prediction set sizes for test data points, after removing duplicate predictions.** Our algorithm constructs adaptive confidence sets that vary in size for different inputs. $\mathbf{T}_{\alpha,\tau}$ is the set of values of tunable parameter T used to generate the initial prediction set before removing duplicates.

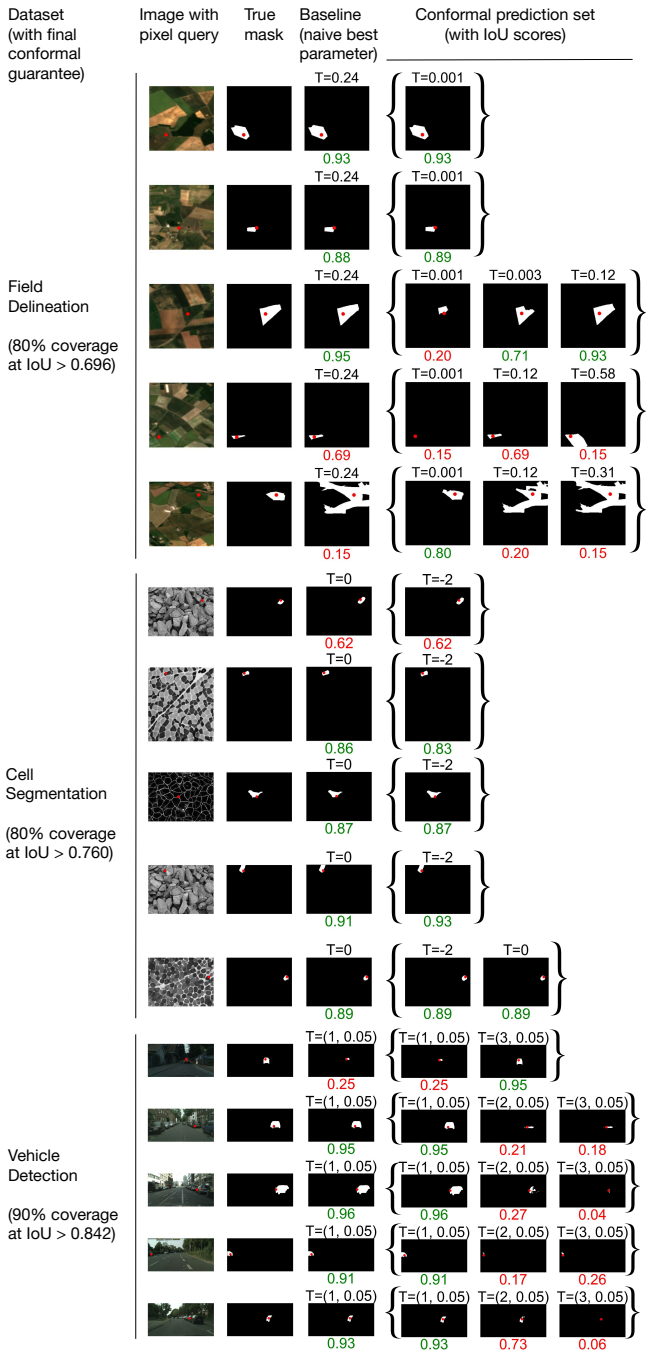

Figure 5: **Conformal prediction sets for example test queries, compared to naive best parameter baseline predictions.** Our conformal sets often include masks of different sizes or shapes, capturing structural uncertainty. Most contain at least one mask with high IoU (green) with the ground truth, and the number of predictions adapts to query difficulty. In contrast, the naive best parameter baseline only outputs a single prediction and does not guarantee high coverage.

single mask.

4 DISCUSSION

We introduced a conformal prediction algorithm for instance segmentation that constructs adaptive sets of candidate masks with provable coverage guarantees. These guarantees formalize the reliability of instance segmentation models by making explicit the probability that at least one mask in the set achieves a desired IoU with the ground truth. In both theory and practice, these sets attain the desired coverage and adapt their size to query difficulty.

Conformal prediction can only perform as well as the underlying segmentation models allow. In our experiments, empirical coverage is capped by the accuracy of the base models, which qualitatively fall short of human-level performance. This gap reflects opportunities for improvement through better algorithms and training strategies. At the same time, some degree of uncertainty is unavoidable, since the imagery itself can be ambiguous — e.g., due to low contrast between adjacent fields in satellite images, overlapping cells in microscopy, or blurry vehicles in street scenes. In such cases, even a perfect model should not be expected to commit to a single mask. There is therefore value in prediction diversity, i.e., representing multiple plausible objects when the imagery is ambiguous. Current segmentation models are optimized to minimize segmentation loss, which does not explicitly incentivize producing multiple plausible alternatives when the input is ambiguous. We leave the development of training strategies that promote greater mask diversity to future work.

Several limitations remain. First, the conformal guarantee is global: every input receives the same IoU threshold, even though some queries are easy while others are ambiguous. Second, our algorithm requires varying a tunable parameter. In settings where such a parameter is not available or does not induce meaningful diversity, constructing conformal sets may be challenging. Lastly, the set construction steps (set cover, duplicate removal) involve combinatorial optimization. Although feasible in our experiments due to small set sizes, these steps may become computationally expensive in some settings. Developing scalable approximations could expand the practical applicability of the method.

Acknowledgements

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0023112. This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any

DMK was partially supported by Generali through its research partnership with the Laboratory for Information and Decision Systems at MIT.

References

Andéol, Léo et al. (2023). "Confident object detection via conformal prediction and conformal risk control: an application to railway signaling." In: *Conformal and Probabilistic Prediction with Applications*. PMLR, pp. 36–55.

Andéol, Léo et al. (2025). "Conformal Object Detection by Sequential Risk Control." In: *arXiv preprint arXiv:2505.24038*.

Angelopoulos, Anastasios N, Rina Foygel Barber, and Stephen Bates (2025a). *Theoretical Foundations of Conformal Prediction*. arXiv: 2411.11824v3 [math.ST]. URL: https://arxiv.org/abs/2411.11824.

Angelopoulos, Anastasios N, Stephen Bates, et al. (2023). "Conformal prediction: A gentle introduction." In: *Foundations and trends® in machine learning* 16.4, pp. 494–591.

Angelopoulos, Anastasios N et al. (2022). "Conformal risk control." In: *arXiv preprint arXiv:2208.02814*.

Angelopoulos, Anastasios N et al. (2025b). "Learn then test: Calibrating predictive algorithms to achieve risk control." In: *The Annals of Applied Statistics* 19.2, pp. 1641–1662.

Bates, Stephen et al. (2021). "Distribution-free, risk-controlling prediction sets." In: *Journal of the ACM (JACM)* 68.6, pp. 1–34.

Blot, Vincent et al. (2024). "Automatically adaptive conformal risk control." In: *arXiv preprint arXiv:2406.17819*.

Chvatal, Vasek (1979). "A greedy heuristic for the set-covering problem." In: *Mathematics of operations research* 4.3, pp. 233–235.

Cordts, Marius et al. (2016). "The Cityscapes dataset for semantic urban scene understanding." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223.

Davenport, Samuel (2024). "Conformal confidence sets for biomedical image segmentation." In: *arXiv preprint arXiv:2410.03406*.

De Grancey, Florence et al. (2022). "Object detection with probabilistic guarantees: A conformal prediction approach." In: *International Conference on Computer Safety, Reliability, and Security*. Springer, pp. 316–329.

Fisch, Adam et al. (2022). "Conformal prediction sets with limited false positives." In: *International Conference on Machine Learning*. PMLR, pp. 6514–6532.

Garey, Michael R. and David S. Johnson (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness*. USA: W. H. Freeman & Co. ISBN: 0716710447.

Johnson, David S (1973). "Approximation algorithms for combinatorial problems." In: *Proceedings of the fifth annual ACM symposium on Theory of computing*, pp. 38–49.

Karp, Richard M. (1972). "Reducibility among Combinatorial Problems." In: *Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations*. Ed. by Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger. Boston, MA: Springer US, pp. 85–103. ISBN: 978-1-4684-2001-2. DOI: 10.1007/978-1-4684-2001-2_9. URL: https://doi.org/10.1007/978-1-4684-2001-2_9.

Kerner, Hannah et al. (2025). "Fields of the world: A machine learning benchmark dataset for global agricultural field boundary segmentation." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 27, pp. 28151–28159.

Kirillov, Alexander et al. (2023). "Segment anything." In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026.

Li, Shuo et al. (2022). "Towards pac multi-object detection and tracking." In: *arXiv preprint arXiv:2204.07482*.

Lovász, László (1975). "On the ratio of optimal integral and fractional covers." In: *Discrete mathematics* 13.4, pp. 383–390.

Luo, Rui and Zhixin Zhou (2025). "Conditional conformal risk adaptation." In: *arXiv preprint arXiv:2504.07611*.

Luo, Rui et al. (2026). "Enhancing image-conditional coverage in segmentation: Adaptive thresholding via differentiable miscoverage loss." In: *14th International Conference on Learning Representations (ICLR 2026)*.

Mossina, Luca, Joseba Dalmau, and Léo Andéol (2024). "Conformal semantic image segmentation: Post-hoc quantification of predictive uncertainty." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3574–3584.

Mossina, Luca and Corentin Friedrich (2025). "Conformal prediction for image segmentation using morphological prediction sets." In: *International Conference on Medical*

Image Computing and Computer-Assisted Intervention. Springer, pp. 78–88.

Mukama, Bruce Cyusa et al. (2024). "Copula-based conformal prediction for object detection: a more efficient approach." In: *13th Symposium on Conformal and Probabilistic Prediction with Applications*. Vol. 230, pp. 140–157.

Pachitariu, Marius, Michael Rariden, and Carsen Stringer (2025). "Cellpose-SAM: superhuman generalization for cellular segmentation." In: *bioRxiv*, pp. 2025–04.

Persello, Claudio and Lorenzo Bruzzone (2009). "A novel protocol for accuracy assessment in classification of very high resolution images." In: *IEEE transactions on geoscience and remote sensing* 48.3, pp. 1232–1244.

Stringer, Carsen et al. (2021). "Cellpose: a generalist algorithm for cellular segmentation." In: *Nature methods* 18.1, pp. 100–106.

Timans, Alexander et al. (2024). "Adaptive bounding box uncertainties via two-step conformal prediction." In: *European Conference on Computer Vision*. Springer, pp. 363–398.

Viti, Bruno, Elias Karabelas, and Martin Holler (2025). "CONSIGN: Conformal Segmentation Informed by Spatial Groupings via Decomposition." In: *arXiv preprint arXiv:2505.14113*.

Waldner, François et al. (2021). "Detect, consolidate, delineate: Scalable mapping of field boundaries using satellite images." In: *Remote sensing* 13.11, p. 2197.

Zouzou, Alya, Mélanie Ducoffe, Ryma Boumazouza, et al. (2025). "Robust vision-based runway detection through conformal prediction and conformal map." In: *arXiv preprint arXiv:2505.16740*.

Supplementary Material

Kerri Lu[1,2] Dan M. Kluger[4] Stephen Bates[1,2] Sherrie Wang[1,3,4]

[1]Laboratory for Information and Decision Systems, MIT, Cambridge, MA, USA
[2]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA
[3]Department of Mechanical Engineering, MIT, Cambridge, MA, USA
[4]Institute for Data, Systems, and Society, MIT, Cambridge, MA, USA

A CODE AND DATA

Code and data are available at https://github.com/conformal-instance-segmentation/conformal-instance-segmentation-code

B LICENSE INFORMATION OF DATASETS

The Fields of the World dataset (Kerner et al., 2025) for France is under the French Open License (Licence Ouverte 2.0). This means the data must be attributed to Registre Parcellaire Graphique (RPG) and its provider, users must avoid implying endorsement, and users may freely use, modify, and redistribute the data for any purpose.

The Cellpose dataset (Stringer et al., 2021) is under the Creative Commons NonCommercial (CC-BY-NC) license.

The Cityscapes dataset (Cordts et al., 2016) is under a custom license: "This Cityscapes Dataset is made freely available to academic and non-academic entities for non-commercial purposes such as academic research, teaching, scientific publications, or personal experimentation. Permission is granted to use the data given that you agree to our license terms." https://www.cityscapes-dataset.com/license

C PROOF OF VALIDITY OF ALGORITHMS

C.1 ADDITIONAL ASSUMPTIONS NEEDED

Recall Assumption 1 that for each n, the calibration samples $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} \mathbb{P}$ and independently, $(X^{\text{test}}, Y^{\text{test}}) \sim \mathbb{P}$. Throughout this supplement, we let (X, Y) denote an arbitrary random object from the distribution $\mathbb{P}$.

The next assumption ensures that $1 - \alpha$ and τ are small enough to be feasible.

Assumption A.2. $\mathbb{P}\Big(\max_{j \in [k]} \{\text{IoU}(Y, f(X, t_j))\} > \tau \Big) > 1 - \alpha.$

Assumption A.2 can be checked empirically, and assuming the calibration sample size n is relatively large, Algorithm 1 will return an error when it does not hold. If Assumption A.2 does not hold, the user can either increase the size of the parameter search space $\{t_1, \ldots, t_k\}$, can increase α, or can decrease τ. They can also consider some combination of these options. Algorithm 1 can subsequently be applied after the parameters are changed such that it will not return an error.

The next assumption is an implementation criterion that the user can enforce. We suspect that the assumption will also hold in most default implementations of Algorithms 1 and 2.

Assumption A.3. In Algorithms 1 and 2, whenever a search across sets of the same cardinality is conducted, the ordering of sets with the same cardinality is fixed. In particular, in Line 8 of Algorithm 1, for each $l \in [k]$, there is a predetermined ordering of all subsets $J \subseteq [k]$ with $|J| = l$. Thus in cases where the $\arg\min$ in Line 8 of Algorithm 1 is equal to a collection $\{J_1, \ldots, J_r\}$ of distinct subsets of $[k]$ with cardinality l, $J_{\alpha,\tau}$ will be set to the first element of $\{J_1, \ldots, J_r\}$ in

the predetermined ordering. Moreover, in Line 2 of Algorithm 2, the for loop always uses the same ordering to search through sets with the same cardinality.

We introduce one more technical assumption that ensures that asymptotically, $J_{\alpha,\tau}$ returned by Algorithm 1 will equal $J^*_{\alpha,\tau}$ with probability converging to 1. First we must introduce some additional notation. For each $J \subseteq [k]$ define

$$\mu_\tau(J) \equiv \mathbb{P}\Big(\max_{j\in J}\big\{\text{IoU}\big(Y, f(X,t_j)\big)\big\} > \tau\Big). \tag{1}$$

Further define

$$\mathcal{A}_{\alpha,\tau} \equiv \big\{J \subseteq [k] \;:\; \mu_\tau(J) \geq 1-\alpha\big\},$$

and observe that, by Assumption A.2, $\mathcal{A}_{\alpha,\tau}$ is nonempty. Now let $J^*_{\alpha,\tau}$ be an element of $\arg\min_{J\subseteq\mathcal{A}_{\alpha,\tau}}\{|J|\}$. Under Assumption A.3, we can further define $J^*_{\alpha,\tau}$ to be the element of $\mathcal{A}_{\alpha,\tau}$ that is first among all elements in $\mathcal{A}_{\alpha,\tau}$ with cardinality $|J^*_{\alpha,\tau}|$ according to the ordering used in Line 8 of Algorithm 1. While by definition, $\mu_\tau(J^*_{\alpha,\tau}) \geq 1-\alpha$, the next assumption states that this inequality is strict, which should generally hold (for example, if there were a continuous prior on all $\mu_\tau(J)$ values, this next assumption would hold with probability 1 under Assumption A.2).

Assumption A.4. $\mathcal{A}_{\alpha,\tau}$ is nonempty and $\mu_\tau(J^*_{\alpha,\tau}) > 1-\alpha$.

C.2 HELPFUL LEMMA

To prove the asymptotic validity of Algorithms 1 and 2, it helps to introduce a lemma establishing the asymptotic behavior of the set $J_{\alpha,\tau}$ returned in Algorithm 1. To this it is convenient to define,

$$\mathcal{A}^c_{\alpha,\tau} \equiv \big\{J \subseteq [k] \;:\; \mu_\tau(J) < 1-\alpha\big\}.$$

Lemma C.1. *Let $J^{(n)}_{\alpha,\tau}$ denote the set of indices returned when running Algorithm 1 with calibration samples $\big((X_i,Y_i)\big)_{i=1}^n$ and parameters $\alpha,\tau \in (0,1)$, and suppose $J^{(n)}_{\alpha,\tau} = \text{NA}$ if Algorithm 1 returns an error. Then under Assumptions 1 and A.2, $\lim_{n\to\infty}\mathbb{P}(J^{(n)}_{\alpha,\tau} \in \mathcal{A}^c_{\alpha,\tau}) = 0$ and $\lim_{n\to\infty}\mathbb{P}(J^{(n)}_{\alpha,\tau} = \text{NA}) = 0$.*

Proof. Fix $\alpha,\tau \in (0,1)$ to be the tolerance and IoU thresholds used in Algorithm 1. Let $\mathcal{X} = \mathbb{R}^{W\times H\times C} \times [W] \times [H]$ and $\mathcal{Y} = \{0,1\}^{W\times H}$ denote the input and output space for $f(\cdot,t)$, respectively. Now for each $J \subseteq [k]$, define $g_J : \mathcal{X}\times\mathcal{Y} \to \{0,1\}$ by

$$g_J(x,y) \equiv \begin{cases} 1 & \text{if IoU}\big(y, f(x,t_j)\big) > \tau \text{ for some } j \in J, \\ 0 & \text{otherwise.}\end{cases}$$

Observe that for all $J \subseteq [k]$,

$$\mathbb{E}[g_J(X,Y)] = \mathbb{P}\Big(\max_{j\in J}\big\{\text{IoU}\big(Y, f(X,t_j)\big)\big\} > \tau\Big) = \mu_\tau(J),$$

where the last step follows by definition of $\mu_\tau(J)$ in Equation (1).

Now fix $J \in \mathcal{A}^c_{\alpha,\tau}$, and we will show that $\lim_{n\to\infty}\mathbb{P}(J^{(n)}_{\alpha,\tau} = J) = 0$. Since, $J \in \mathcal{A}^c_{\alpha,\tau}$, $\mathbb{E}[g_J(X,Y)] = \mu_\tau(J) < 1-\alpha$. Further by by Assumption 1, $g_J(X_i,Y_i) \in \{0,1\}$ are IID for $i = 1,\ldots,n$, so by the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^n g_J(X_i,Y_i) \xrightarrow{p} \mu_\tau(J) < 1-\alpha \quad \text{as} \quad n\to\infty. \tag{2}$$

Further observe that by Algorithm 1, if the returned set of indices $J^{(n)}_{\alpha,\tau}$ equals J, it must be the case that $\sum_{i=1}^n g_J(X_i,Y_i) \geq (1-\alpha)n$. Hence

$$\mathbb{P}(J^{(n)}_{\alpha,\tau} = J) \leq \mathbb{P}\Big(\sum_{i=1}^n g_J(X_i,Y_i) \geq (1-\alpha)n\Big) \leq \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n g_J(X_i,Y_i) - \mu_\tau(J)\Big| \geq (1-\alpha) - \mu_\tau(J)\Big).$$

Taking the limit as $n \to \infty$ of each side, by (2) and the definition of convergence in probability,

$$\limsup_{n\to\infty} \mathbb{P}(J_{\alpha,\tau}^{(n)} = J) \leq \limsup_{n\to\infty} \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} g_J(X_i, Y_i) - \mu_\tau(J)\Big| \geq (1-\alpha) - \mu_\tau(J)\Big) = 0.$$

Hence we have shown that $\limsup_{n\to\infty} \mathbb{P}(J_{\alpha,\tau}^{(n)} = J) = 0$, and this argument holds for any fixed $J \in \mathcal{A}_{\alpha,\tau}^c$.

By the union bound and applying this result,

$$0 \leq \limsup_{n\to\infty} \mathbb{P}(J_{\alpha,\tau}^{(n)} \in \mathcal{A}_{\alpha,\tau}^c) = \limsup_{n\to\infty} \mathbb{P}\Big(\bigcup_{J\in\mathcal{A}_{\alpha,\tau}^c} \{J_{\alpha,\tau}^{(n)} = J\}\Big) \leq \limsup_{n\to\infty} \sum_{J\in\mathcal{A}_{\alpha,\tau}^c} \mathbb{P}(J_{\alpha,\tau}^{(n)} = J) = 0.$$

The above result further implies that $\lim_{n\to\infty} \mathbb{P}(J_{\alpha,\tau}^{(n)} \in \mathcal{A}_{\alpha,\tau}^c) = 0$.

To complete the proof we must show that $\lim_{n\to\infty} \mathbb{P}(J_{\alpha,\tau}^{(n)} = \text{NA}) = 0$. By Assumption A.2,

$$\mathbb{E}[g_{[k]}(X, Y)] = \mu_\tau([k]) = \mathbb{P}\Big(\max_{j\in[k]}\{\text{IoU}\big(Y, f(X, t_j)\big)\} > \tau\Big) > 1 - \alpha.$$

Since by Assumption 1, $g_{[k]}(X_i, Y_i) \in \{0, 1\}$ are IID for $i = 1, \ldots, n$, by the weak law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n} g_{[k]}(X_i, Y_i) \xrightarrow{p} \mu_\tau([k]) > 1 - \alpha \quad \text{as} \quad n \to \infty. \tag{3}$$

Now for each $n \in \mathbb{Z}_+$, note that $J_{\alpha,\tau}^{(n)} = \text{NA}$ if and only if Algorithm 1 returns an error in Line 8. Furthermore Algorithm 1 returns an error in Line 8 if and only if $\sum_{i=1}^{n} g_{[k]}(X_i, Y_i) < (1-\alpha)n$. Thus

$$\mathbb{P}(J_{\alpha,\tau}^{(n)} = \text{NA}) = \mathbb{P}\Big(\sum_{i=1}^{n} g_{[k]}(X_i, Y_i) < (1-\alpha)n\Big) \leq \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} g_{[k]}(X_i, Y_i) - \mu_\tau([k])\Big| > \mu_\tau([k]) - (1-\alpha)\Big).$$

Taking the limit as $n \to \infty$ of each side, by (3) and the definition of convergence in probability,

$$\limsup_{n\to\infty} \mathbb{P}(J_{\alpha,\tau}^{(n)} = \text{NA}) \leq \limsup_{n\to\infty} \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} g_{[k]}(X_i, Y_i) - \mu_\tau([k])\Big| \geq (1-\alpha) - \mu_\tau(J)\Big) = 0.$$

By nonnegativity of probabilities, $\lim_{n\to\infty} \mathbb{P}(J_{\alpha,\tau}^{(n)} = \text{NA}) = 0$. □

C.3 VALIDITY OF ALGORITHM 1

The next proposition formally states the asymptotic validity of Algorithm 1.

Proposition C.1. *Let $C_{\alpha,\tau}^{(n)}(X^{\text{test}})$ denote the set of segmentation masks returned when running Algorithm 1 with calibration samples $\big((X_i, Y_i)\big)_{i=1}^n$ and parameters $\alpha, \tau \in (0, 1)$. Under Assumptions 1 and A.2,*

$$\liminf_{n\to\infty} \mathbb{P}\Big(\max_{\hat{y}\in C_{\alpha,\tau}^{(n)}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}}, \hat{y})\big\} > \tau\Big) \geq 1 - \alpha.$$

Proof. Fix $\alpha, \tau \in (0, 1)$ to be the tolerance and IoU thresholds used in Algorithm 1. Let $J_{\alpha,\tau}^{(n)}$ denote the set of indices returned when running Algorithm 1 with calibration samples $\big((X_i, Y_i)\big)_{i=1}^n$ and parameters $\alpha, \tau \in (0, 1)$, and let $J_{\alpha,\tau}^{(n)} = \text{NA}$ if Algorithm 1 returns an error. Because either, $J_{\alpha,\tau}^{(n)} \in \mathcal{A}_{\alpha,\tau}$ and $J_{\alpha,\tau}^{(n)} \in \mathcal{A}_{\alpha,\tau}^c$ or $J_{\alpha,\tau}^{(n)} = \text{NA}$ with each of these events being mutually exclusive,

$$\mathbb{P}(J_{\alpha,\tau}^{(n)} \in \mathcal{A}_{\alpha,\tau}) = 1 - \mathbb{P}(J_{\alpha,\tau}^{(n)} \in \mathcal{A}_{\alpha,\tau}^c) - \mathbb{P}(J_{\alpha,\tau}^{(n)} = \text{NA}).$$

By Lemma C.1, $\lim_{n\to\infty} \mathbb{P}(J_{\alpha,\tau}^{(n)} = \text{NA}) = 0$ and $\lim_{n\to\infty} \mathbb{P}(J_{\alpha,\tau}^{(n)} \in \mathcal{A}_{\alpha,\tau}^c) = 0$, so taking the limit as $n \to \infty$ of each side of the above expression yields that $\lim_{n\to\infty} \mathbb{P}(J_{\alpha,\tau}^{(n)} \in \mathcal{A}_{\alpha,\tau}) = 1$.

In Algroithm 1, provided that $J^{(n)}_{\alpha,\tau} \neq \text{NA}$, $C^{(n)}_{\alpha,\tau}(X^{\text{test}}) = \{f(X^{\text{test}}, t_j) \,:\, j \in J^{(n)}_{\alpha,\tau}\}$. Hence,

$$\begin{aligned}
\mathbb{P}\Big(\max_{\hat{y}\in C^{(n)}_{\alpha,\tau}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}}, \hat{y})\big\} > \tau\Big) &\geq \mathbb{P}\Big(\Big\{\max_{\hat{y}\in C^{(n)}_{\alpha,\tau}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}}, \hat{y})\big\} > \tau\Big\} \cap \{J^{(n)}_{\alpha,\tau} \in \mathcal{A}_{\alpha,\tau}\}\Big) \\
&= \mathbb{P}\Big(\Big\{\max_{j\in J^{(n)}_{\alpha,\tau}} \big\{\text{IoU}\big(Y^{\text{test}}, f(X^{\text{test}}, t_j)\big)\big\} > \tau\Big\} \cap \{J^{(n)}_{\alpha,\tau} \in \mathcal{A}_{\alpha,\tau}\}\Big) \\
&= \sum_{J\in\mathcal{A}_{\alpha,\tau}} \mathbb{P}(J^{(n)}_{\alpha,\tau} = J)\mathbb{P}\Big(\max_{j\in J^{(n)}_{\alpha,\tau}} \big\{\text{IoU}\big(Y^{\text{test}}, f(X^{\text{test}}, t_j)\big)\big\} > \tau \Big| J^{(n)}_{\alpha,\tau} = J\Big) \\
&= \sum_{J\in\mathcal{A}_{\alpha,\tau}} \mathbb{P}(J^{(n)}_{\alpha,\tau} = J)\mathbb{P}\Big(\max_{j\in J} \big\{\text{IoU}\big(Y^{\text{test}}, f(X^{\text{test}}, t_j)\big)\big\} > \tau \Big| J^{(n)}_{\alpha,\tau} = J\Big) \\
&= \sum_{J\in\mathcal{A}_{\alpha,\tau}} \mathbb{P}(J^{(n)}_{\alpha,\tau} = J)\mathbb{P}\Big(\max_{j\in J} \big\{\text{IoU}\big(Y^{\text{test}}, f(X^{\text{test}}, t_j)\big)\big\} > \tau\Big) \\
&= \sum_{J\in\mathcal{A}_{\alpha,\tau}} \mathbb{P}(J^{(n)}_{\alpha,\tau} = J)\mu_\tau(J) \\
&\geq (1-\alpha) \sum_{J\in\mathcal{A}_{\alpha,\tau}} \mathbb{P}(J^{(n)}_{\alpha,\tau} = J) \\
&= (1-\alpha)\mathbb{P}(J^{(n)}_{\alpha,\tau} \in \mathcal{A}_{\alpha,\tau}).
\end{aligned}$$

Above the third step follows from the law of total probability, the fifth step holds because $J^{(n)}_{\alpha,\tau}$ is independent of $(X^{\text{test}}, Y^{\text{test}})$ by Assumption 1 (with the former being a function of the calibration sample $\big((X_i, Y_i)\big)_{i=1}^n$), and the sixth and seventh steps following from the definition of $\mu_\tau(J)$ and $\mathcal{A}_{\alpha,\tau}$ respectively. Recalling that $\lim_{n\to\infty} \mathbb{P}(J^{(n)}_{\alpha,\tau} \in \mathcal{A}_{\alpha,\tau}) = 1$, we can take the $\liminf$ as $n \to \infty$ of each side of the above expression to get that

$$\liminf_{n\to\infty} \mathbb{P}\Big(\max_{\hat{y}\in C^{(n)}_{\alpha,\tau}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}}, \hat{y})\big\} > \tau\Big) \geq (1-\alpha) \cdot \liminf_{n\to\infty} \mathbb{P}(J^{(n)}_{\alpha,\tau} \in \mathcal{A}_{\alpha,\tau}) = (1-\alpha).$$

□

C.4 PROOF OF THEOREM 2.1

For convenience we restate Theorem 2.1 from the main text below. Subsequently, a proof is provided.

Let $C^{(n)}_{\alpha,\tau,\eta}(X^{\text{test}})$ and $\tilde{\theta}^{(n)}$ denote the set of segmentation masks and IoU *threshold returned when running Algorithm 2 with calibration samples $\big((X_i, Y_i)\big)_{i=1}^n$ and parameters $\alpha, \tau, \eta \in (0,1)$. Under Assumptions 1, A.3, and A.4,*

$$\liminf_{n\to\infty} \mathbb{P}\Big(\max_{\hat{y}\in C^{(n)}_{\alpha,\tau,\eta}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}}, \hat{y})\big\} \geq \tilde{\theta}^{(n)}\Big) \geq 1-\alpha.$$

Proof. Fix $\alpha, \tau, \eta \in (0,1)$ to be the tolerance and IoU thresholds used in Algorithm 2. Let $J^{(n)}_{\alpha,\tau}$ denote the set of indices returned when running Algorithm 1 (in Line 11 of Algorithm 2) with calibration samples $\big((X_i, Y_i)\big)_{i=1}^n$ and parameters $\alpha, \tau \in (0,1)$, and let $J^{(n)}_{\alpha,\tau} = \text{NA}$ if Algorithm 1 returns an error.

Recall that $J^*_{\alpha,\tau}$ is a fixed element of $\arg\min_{J\subseteq\mathcal{A}_{\alpha,\tau}}\{|J|\}$, which under Assumption A.3 was set to be the element of $\mathcal{A}_{\alpha,\tau}$ that is first among all elements in $\mathcal{A}_{\alpha,\tau}$ with the minimum cardinality, according to the ordering used in Line 8 of Algorithm 1. We will first show that $\lim_{n\to\infty} \mathbb{P}(J^{(n)}_{\alpha,\tau} = J^*_{\alpha,\tau}) = 1$. To do this, let

$$\mathcal{A}^{(n)}_{\alpha,\tau} = \Big\{J \subseteq [k] \,:\, \sum_{i=1}^n \mathbb{1}\big\{\max_{j\in J} \text{IoU}\big(Y_i, f(X_i, t_j)\big) > \tau\big\} \geq (1-\alpha)n\Big\}$$

and note that in Line 8 of Algorithm 1, $J^{(n)}_{\alpha,\tau}$ is set to the smallest element of $\mathcal{A}^{(n)}_{\alpha,\tau}$ (where the preset ordering is used if $\mathcal{A}^{(n)}_{\alpha,\tau}$ has multiple sets of the same cardinality). Because $J^{(n)}_{\alpha,\tau}$ and $J^*_{\alpha,\tau}$ are the smallest sets in $\mathcal{A}^{(n)}_{\alpha,\tau}$ and $\mathcal{A}_{\alpha,\tau}$ (with ties between

sets of the same size broken according to the same ordering), $J^*_{\alpha,\tau} \in \mathcal{A}^{(n)}_{\alpha,\tau}$ implies that either $J^{(n)}_{\alpha,\tau} \notin \mathcal{A}_{\alpha,\tau}$ or $J^{(n)}_{\alpha,\tau} = J^*_{\alpha,\tau}$. Hence if both $J^{(n)}_{\alpha,\tau} \in \mathcal{A}_{\alpha,\tau}$ and $J^*_{\alpha,\tau} \in \mathcal{A}^{(n)}_{\alpha,\tau}$ then $J^{(n)}_{\alpha,\tau} = J^*_{\alpha,\tau}$. Thus by monotonicity of probability measure and the union bound,

$$\begin{aligned}
\mathbb{P}(J^{(n)}_{\alpha,\tau} = J^*_{\alpha,\tau}) &\geq \mathbb{P}\big(\{J^{(n)}_{\alpha,\tau} \in \mathcal{A}_{\alpha,\tau}\} \cap \{J^*_{\alpha,\tau} \in \mathcal{A}^{(n)}_{\alpha,\tau}\}\big) \\
&= 1 - \mathbb{P}\big(\{J^{(n)}_{\alpha,\tau} \in \mathcal{A}_{\alpha,\tau}\}^c \cup \{J^*_{\alpha,\tau} \in \mathcal{A}^{(n)}_{\alpha,\tau}\}^c\big) \\
&\geq 1 - \mathbb{P}\big(\{J^{(n)}_{\alpha,\tau} \in \mathcal{A}_{\alpha,\tau}\}^c\big) - \mathbb{P}\big(\{J^*_{\alpha,\tau} \in \mathcal{A}^{(n)}_{\alpha,\tau}\}^c\big) \\
&= 1 - \mathbb{P}\big(J^{(n)}_{\alpha,\tau} \in \mathcal{A}^c_{\alpha,\tau}\big) - \mathbb{P}\big(J^{(n)}_{\alpha,\tau} = \text{NA}\big) - \mathbb{P}\big(\{J^*_{\alpha,\tau} \in \mathcal{A}^{(n)}_{\alpha,\tau}\}^c\big).
\end{aligned}$$

Let $\mathcal{X} = \mathbb{R}^{W\times H\times C} \times [W] \times [H]$ and $\mathcal{Y} = \{0,1\}^{W\times H}$ denote the input and output space for $f(\cdot, t)$, respectively, and define $h : \mathcal{X} \times \mathcal{Y} \to \{0,1\}$ by

$$h(x,y) \equiv \begin{cases} 1 & \text{if IoU}\big(y, f(x,t_j)\big) > \tau \text{ for some } j \in J^*_{\alpha,\tau}, \\ 0 & \text{otherwise.} \end{cases}$$

Observe that by Assumption A.4,

$$\mathbb{E}[h(X,Y)] = \mathbb{P}\Big(\max_{j\in J^*_{\alpha,\tau}} \big\{\text{IoU}\big(Y, f(X,t_j)\big)\big\} > \tau\Big) = \mu_\tau(J^*_{\alpha,\tau}) > 1-\alpha,$$

where the penultimate step follows by definition of $\mu_\tau(J)$ in Equation (1) and the last step follows from Assumption A.4. Since by Assumption 1, $h(X_i, Y_i) \in \{0,1\}$ are IID for $i = 1,\dots,n$, by the weak law of large numbers

$$\frac{1}{n}\sum_{i=1}^n h(X_i,Y_i) \xrightarrow{p} \mu_\tau(J^*_{\alpha,\tau}) > 1-\alpha \quad \text{as} \quad n\to\infty. \tag{4}$$

Since, by definition of $h(\cdot,\cdot)$ and $\mathcal{A}^{(n)}_{\alpha,\tau}$, $\sum_{i=1}^n h(X_i,Y_i) \geq (1-\alpha)n$ if and only if $J^*_{\alpha,\tau} \in \mathcal{A}^{(n)}_{\alpha,\tau}$,

$$\mathbb{P}\big(\{J^*_{\alpha,\tau} \in \mathcal{A}^{(n)}_{\alpha,\tau}\}^c\big) = \mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^n h(X_i,Y_i) < (1-\alpha)\Big) \leq \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n h(X_i,Y_i) - \mu_\tau(J^*_{\alpha,\tau})\Big| \geq \mu_\tau(J^*_{\alpha,\tau}) - (1-\alpha)\Big).$$

Taking the limit as $n\to\infty$ of each side, by (4) and the definition of convergence in probability,

$$\limsup_{n\to\infty} \mathbb{P}\big(\{J^*_{\alpha,\tau} \in \mathcal{A}^{(n)}_{\alpha,\tau}\}^c\big) \leq \limsup_{n\to\infty} \mathbb{P}\Big(\Big|\frac{1}{n}\sum_{i=1}^n h(X_i,Y_i) - \mu_\tau(J^*_{\alpha,\tau})\Big| \geq \mu_\tau(J^*_{\alpha,\tau}) - (1-\alpha)\Big) = 0.$$

By nonnegativity of probabilities $\lim_{n\to\infty} \mathbb{P}\big(\{J^*_{\alpha,\tau} \in \mathcal{A}^{(n)}_{\alpha,\tau}\}^c\big) = 0$. Also note that since Assumption A.4 implies Assumption A.2, by Lemma C.1, $\lim_{n\to\infty} \mathbb{P}(J^{(n)}_{\alpha,\tau} = \text{NA}) = 0$ and $\lim_{n\to\infty} \mathbb{P}(J^{(n)}_{\alpha,\tau} \in \mathcal{A}^c_{\alpha,\tau}) = 0$. Combining these results with an earlier inequality,

$$\liminf_{n\to\infty} \mathbb{P}(J^{(n)}_{\alpha,\tau} = J^*_{\alpha,\tau}) \geq 1 - \limsup_{n\to\infty} \mathbb{P}\big(J^{(n)}_{\alpha,\tau} \in \mathcal{A}^c_{\alpha,\tau}\big) - \limsup_{n\to\infty} \mathbb{P}\big(J^{(n)}_{\alpha,\tau} = \text{NA}\big) - \limsup_{n\to\infty} \mathbb{P}\big(\{J^*_{\alpha,\tau} \in \mathcal{A}^{(n)}_{\alpha,\tau}\}^c\big) = 1.$$

Since probabilities are upper bounded by 1, $\lim_{n\to\infty} \mathbb{P}(J^{(n)}_{\alpha,\tau} = J^*_{\alpha,\tau}) = 1$.

We are now ready to investigate the limiting behavior of Algorithm 2. To do this, let $\mathcal{P}_k$ denote the collection of subsets of $[k]$. Observe that,

$$\mathbb{P}\Big(\max_{\hat{y}\in C^{(n)}_{\alpha,\tau,\eta}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}},\hat{y})\big\} \geq \tilde{\theta}^{(n)}\Big) = \sum_{J\in\mathcal{P}_k\cup\{\text{NA}\}} \mathbb{P}\Big(\Big\{\max_{\hat{y}\in C^{(n)}_{\alpha,\tau,\eta}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}},\hat{y})\big\} \geq \tilde{\theta}^{(n)}\Big\} \cap \{J^{(n)}_{\alpha,\tau} = J\}\Big).$$

Taking the limit as $n\to\infty$ of each side and noting that for $J \neq J^*_{\alpha,\tau}$, $\lim_{n\to\infty} \mathbb{P}(J^{(n)}_{\alpha,\tau} = J) = 0$ (because $\lim_{n\to\infty} \mathbb{P}(J^{(n)}_{\alpha,\tau} = J^*_{\alpha,\tau}) = 1$), and hence $\lim_{n\to\infty} \mathbb{P}\big(E_n \cap \{J^{(n)}_{\alpha,\tau} = J\}\big) = 0$ for any sequence of events E_n and $J \neq J^*_{\alpha,\tau}$, it follows that

$$\lim_{n\to\infty} \mathbb{P}\Big(\max_{\hat{y}\in C^{(n)}_{\alpha,\tau,\eta}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}},\hat{y})\big\} \geq \tilde{\theta}^{(n)}\Big) = \lim_{n\to\infty} \mathbb{P}\Big(\Big\{\max_{\hat{y}\in C^{(n)}_{\alpha,\tau,\eta}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}},\hat{y})\big\} \geq \tilde{\theta}^{(n)}\Big\} \cap \{J^{(n)}_{\alpha,\tau} = J^*_{\alpha,\tau}\}\Big).$$

To simplify the above expression let $\mathcal{B}$ denote the set of all possible collections of binary masks in $\mathcal{Y}$. Define $u : \mathcal{X} \times \mathcal{P}_k \to \mathcal{B}$ be the function such that

$$u(x, J) \equiv \texttt{UNIQUE}\big(\{f(x, t_j) \,:\, j \in J\}, \eta\big) \quad \text{for all } x \in \mathcal{X} \quad \text{and} \quad J \subseteq [k],$$

where UNIQUE is the procedure defined in Algorithm 2. Note that by Assumption A.3, the procedure UNIQUE is deterministic and hence $u(\cdot, \cdot)$ is a deterministic function. Next, define $M : \mathcal{X} \times \mathcal{Y} \times \mathcal{P}_k \to [0, 1]$ to be the function given by

$$M(x, y, J) \equiv \max_{\hat{y} \in u(x,J)} \big\{\text{IoU}(y, \hat{y})\big\} \quad \text{for all } x \in \mathcal{X}, \quad y \in \mathcal{Y}, \quad \text{and} \quad J \subseteq [k].$$

Note that $M : \mathcal{X} \times \mathcal{Y} \times \mathcal{P}_k \to [0, 1]$ is also a deterministic function. Finally let $\mathcal{Q}_\alpha(\cdot)$ be a deterministic operator that takes in a sequence of real numbers as input and returns the α empirical quantile (as defined in Definition 2.9 of Angelopoulos et al. (2025a)). Now observe that by Algorithm 2

$$\max_{\hat{y} \in C^{(n)}_{\alpha,\tau,\eta}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}}, \hat{y})\big\} = M(X^{\text{test}}, Y^{\text{test}}, J^{(n)}_{\alpha,\tau}) \quad \text{and} \quad \tilde{\theta}^{(n)} = \mathcal{Q}_\alpha\Big(\big\{M(X_i, Y_i, J^{(n)}_{\alpha,\tau})\big\}_{i=1}^n\Big),$$

provided that $J^{(n)}_{\alpha,\tau} \neq \text{NA}$. Hence letting $o(1)$ denote any term that converges to 0 as $n \to \infty$ and letting B_n and E_n denote the following events

$$B_n \equiv \Big\{M(X^{\text{test}}, Y^{\text{test}}, J^*_{\alpha,\tau}) \geq \mathcal{Q}_\alpha\Big(\big\{M(X_i, Y_i, J^*_{\alpha,\tau})\big\}_{i=1}^n\Big)\Big\} \quad \text{and} \quad E_n \equiv \{J^{(n)}_{\alpha,\tau} = J^*_{\alpha,\tau}\},$$

an earlier result simplifies to

$$\begin{aligned}
\mathbb{P}\Big(\max_{\hat{y} \in C^{(n)}_{\alpha,\tau,\eta}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}}, \hat{y})\big\} \geq \tilde{\theta}^{(n)}\Big) &= \mathbb{P}\Big(\Big\{\max_{\hat{y} \in C^{(n)}_{\alpha,\tau,\eta}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}}, \hat{y})\big\} \geq \tilde{\theta}^{(n)}\Big\} \cap \{J^{(n)}_{\alpha,\tau} = J^*_{\alpha,\tau}\}\Big) + o(1) \\
&= \mathbb{P}\Big(\Big\{M(X^{\text{test}}, Y^{\text{test}}, J^{(n)}_{\alpha,\tau}) \geq \mathcal{Q}_\alpha\Big(\big\{M(X_i, Y_i, J^{(n)}_{\alpha,\tau})\big\}_{i=1}^n\Big)\Big\} \cap E_n\Big) + o(1) \\
&= \mathbb{P}\Big(\Big\{M(X^{\text{test}}, Y^{\text{test}}, J^*_{\alpha,\tau}) \geq \mathcal{Q}_\alpha\Big(\big\{M(X_i, Y_i, J^*_{\alpha,\tau})\big\}_{i=1}^n\Big)\Big\} \cap E_n\Big) + o(1) \\
&= \mathbb{P}(B_n \cap E_n) + o(1).
\end{aligned}$$

Above the penultimate step holds because $J^{(n)}_{\alpha,\tau} = J^*_{\alpha,\tau}$ under the event E_n. Now we will show that for each $n \in \mathbb{Z}_+$, $\mathbb{P}(B^c_n) < \alpha$. To do this fix $n \in \mathbb{Z}_+$, and define $\xi_{n+1} \equiv M(X^{\text{test}}, Y^{\text{test}}, J^*_{\alpha,\tau})$, and $\xi_i \equiv M(X_i, Y_i, J^*_{\alpha,\tau})$ for each $i = 1, \ldots, n$. Since $J^*_{\alpha,\tau}$ is a fixed set of indices and by Assumption 1, $\big((X_i, Y_i)\big)_{i=1}^n \overset{\text{iid}}{\sim} \mathbb{P}$ and independently $(X^{\text{test}}, Y^{\text{test}}) \sim \mathbb{P}$, it follows that $(\xi_i)_{i=1}^{n+1}$ is an exchangeable sequence of random variables. Note that

$$\mathbb{P}(B^c_n) = \mathbb{P}\Big(M(X^{\text{test}}, Y^{\text{test}}, J^*_{\alpha,\tau}) < \mathcal{Q}_\alpha\Big(\big\{M(X_i, Y_i, J^*_{\alpha,\tau})\big\}_{i=1}^n\Big)\Big) = \mathbb{P}\Big(\xi_{n+1} < \mathcal{Q}_\alpha\big(\{\xi_i\}_{i=1}^n\big)\Big) < \alpha.$$

Above the last step follows from Fact 2.15(ii) in (Angelopoulos et al., 2025a) because $(\xi_i)_{i=1}^{n+1}$ is an exchangeable sequence of random variables. Hence we have shown that $\mathbb{P}(B^c_n) < \alpha$ and this argument holds for any $n \in \mathbb{Z}_+$. Since probability measures are upper bounded by 1 we can apply this result to get that that for all $n \in \mathbb{Z}_+$,

$$\mathbb{P}(B_n \cap E_n) = \mathbb{P}(B_n) + \mathbb{P}(E_n) - \mathbb{P}(B_n \cup E_n) \geq \mathbb{P}(B_n) + \mathbb{P}(E_n) - 1 = \mathbb{P}(E_n) - \mathbb{P}(B^c_n) > \mathbb{P}(E_n) - \alpha.$$

Since we have shown that $\lim_{n\to\infty} \mathbb{P}(E_n) = \lim_{n\to\infty} \mathbb{P}(J^{(n)}_{\alpha,\tau} = J^*_{\alpha,\tau}) = 1$, we can take the limit as $n \to \infty$ of each side of the above inequality to get that $\liminf_{n\to\infty} \mathbb{P}(B_n \cap E_n) \geq 1 - \alpha$. Combining this with an earlier result we get that

$$\liminf_{n\to\infty} \mathbb{P}\Big(\max_{\hat{y} \in C^{(n)}_{\alpha,\tau,\eta}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}}, \hat{y})\big\} \geq \tilde{\theta}^{(n)}\Big) = \liminf_{n\to\infty} \Big(\mathbb{P}(B_n \cap E_n) + o(1)\Big) = \liminf_{n\to\infty} \mathbb{P}(B_n \cap E_n) \geq 1 - \alpha.$$

□

D IOU HISTOGRAMS FOR CALIBRATION DATA

Our conformal algorithm requires the selection of a target error rate α and IoU threshold τ. To guide our choices, we plotted histograms of the maximum IoU achievable for each calibration data point under any value of the tunable parameter T (Figure D.1). For each experiment, we minimized the target error rate α while maximizing IoU threshold τ, subject to feasibility on the calibration dataset.

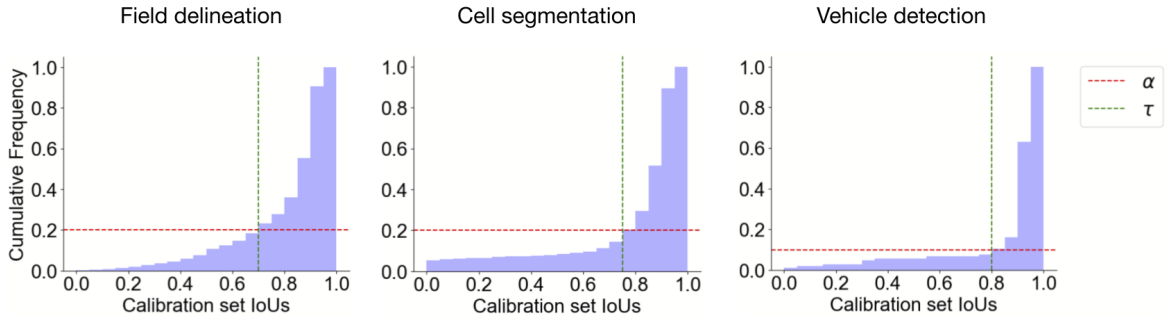

Figure D.1: **Cumulative distributions of maximum IoUs over all T values, for calibration data points.** For each experiment, we minimized the target error rate α while maximizing IoU threshold τ, subject to feasibility (i.e., there must be less than α fraction of calibration points with IoU $< \tau$).

E WHY THE NAIVE BEST PARAMETER BASELINE HAS LOW COVERAGE

While our conformal method outputs a confidence set of multiple parameter values, the naive best parameter baseline outputs the single parameter value that results in the highest coverage on the calibration dataset. The disadvantage of the naive best parameter method is that there does not always exist a single parameter value that results in accurate predictions for most calibration points. In particular, in our instance segmentation setting, there does not always exist a single parameter value that results in mask predictions with high IoU for $(1 - \alpha)$ fraction of the calibration points. Thus, the naive best parameter baseline often cannot guarantee high coverage on the test set.

As an illustrative example (Figure E.1), we take the field delineation dataset and plot the fraction of calibration predictions with IoU greater than $\tau = 0.7$ at each single parameter value T (which ranges from 0 to 1). The target coverage for the field delineation example in our paper was $(1 - \alpha) = 0.8$, shown as a red line in the figure. We see that there is no single parameter value that attains the target coverage of 80% on the calibration set; the best parameter value ($T = 0.24$) results in about 67% coverage. Since the naive best parameter baseline only outputs a single parameter value, it will fail to reach the target coverage on the calibration set and is thus unlikely to reach the target coverage on the test set. In contrast, our conformal method constructs a set of multiple parameter values such that for about 80% of the test points, at least one value will result in high IoU.

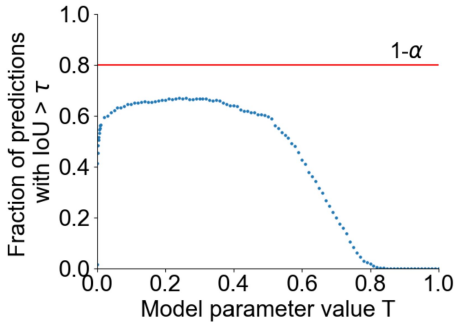

Figure E.1: **Fraction of calibration predictions with IoU greater than $\tau = 0.7$ at each model parameter value T for the field delineation dataset.** No single parameter value attains the target coverage of 80% on the calibration set.

F COMPARISON TO MORPHOLOGICAL DILATION BASELINE

We compare our method to the morphological dilation-based method from Mossina and Friedrich (2025). This method constructs conformal confidence sets for binary segmentation using dilation, i.e., adding a margin to a predicted mask, where the margin size is a fixed number determined using calibration data. This algorithm only guarantees that the true mask is contained within the dilated mask with high probability, and does not guarantee high IoU.

Experimental setup For the comparison, we use our field delineation dataset. For each sampled pixel, we first generate a single mask prediction at the best parameter value as defined in the previous paragraph, which is $T = 0.24$. For each calibration pixel, we dilate the predicted mask by D pixels, for all values of D from 0 to 30, and record the smallest D value where the dilated mask contains the true field. Finally we compute the 80th percentile of these minimum D values, which is $D = 2$. The conformal guarantee is that for a test pixel, with 80% probability, dilating the predicted mask by $D = 2$ pixels will result in a mask that contains the true field.

Results The dilation-based method results in low coverage for IoU, undersegmentation, and lack of flexibility in predictions. We visualize results for five example test points (Figure F.1). Three of the five dilated masks have IoU less than 0.7. Overall, only 54.4% of the 917 test points have dilated masks with IoU greater than 0.7. This low coverage is expected as the dilation-based method does not guarantee high IoUs, as discussed above. Our paper's method achieves 79.7% coverage. The dilated masks tend to undersegment the fields because the dilation-based method is designed to output a single mask that contains the true field with high probability (see Appendix G for a quantitative comparison of the undersegmentation against our method). Finally, each dilated mask is constrained to be a modification of a single prediction, while our method's conformal sets can contain a diverse range of possible fields from multiple predictions, allowing users to select from multiple plausible options.

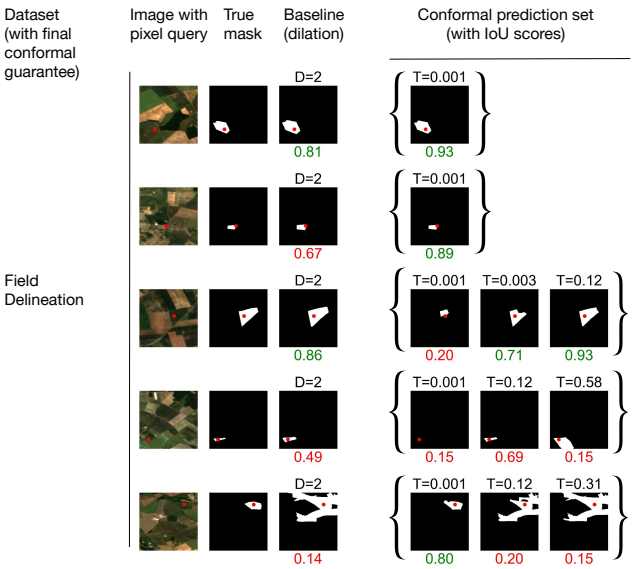

Figure F.1: **Conformal prediction sets for example test queries, compared to dilation baseline predictions.** The third column shows the mask prediction for each test point dilated by $D = 2$ pixels, while the rightmost column shows the conformal prediction sets generated by our method. While most of our conformal sets contain at least one mask with high IoU (green) with the ground truth, the dilation-based method does not provide guarantees on IoU.

G QUANTIFYING STRUCTURAL UNCERTAINTY

We are interested in capturing structural uncertainty in the size and shape of objects, including over- and under-segmentation: in applications such as field delineation, there is often uncertainty about whether adjacent regions should be treated as a single object or split into multiple objects. Qualitatively we find that our prediction sets contain masks of diverse sizes and shapes, while previous works only predict a single mask or modifications of a single baseline mask (e.g., dilation or boundary expansion).

To quantify our method's performance, we will use the oversegmentation score S_{over} and undersegmentation score S_{under} from Persello and Bruzzone (2009). For a true mask Y and predicted mask $\hat{Y}$, they are defined as follows:

$$S_{\text{over}} = 1 - |Y \cap \hat{Y}|/|Y|$$
$$S_{\text{under}} = 1 - |Y \cap \hat{Y}|/|\hat{Y}|$$

where the absolute value sign denotes the area of a mask. The scores range from 0 to 1, where 0 is a perfect match and higher scores correspond to more over- and under-segmentation respectively.

For each test point in the field delineation dataset, we compute these scores between the true mask and each of the masks in our conformal prediction set. We record the minimum oversegmentation and undersegmentation score over the masks in each prediction set. We also compute these scores for the dilation-based conformal baseline (Appendix F). We plot histograms of the scores, comparing our method with the baseline (Figure G.1). We find that both methods have low oversegmentation, but the baseline method suffers from undersegmentation compared to our method. This is because the dilation-based method is designed to output a conformal mask that contains the true field, resulting in field predictions that are too large. In contrast, our method is designed to guarantee high IoU by outputting a set of multiple predictions, so that there is likely at least one prediction that avoids undersegmentation (as well as at least one prediction that avoids oversegmentation).

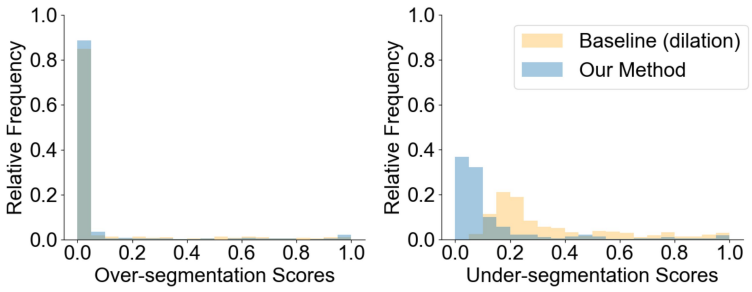

Figure G.1: **Over- and under-segmentation scores for our method and the dilation-based conformal baseline, over the field delineation test set.** Both methods have low oversegmentation, but the baseline method suffers from undersegmentation.

H MODIFIED CONFORMAL ALGORITHM WITH FINITE SAMPLE GUARANTEES

Algorithm 3 Compute conformal prediction set for instance segmentation, with finite sample guarantees

Input:

- calibration data, randomly split into two sets of size n_1 and n_2: $\{(X_1, Y_1), \ldots, (X_{n_1}, Y_{n_1})\}$ and $\{(X_{n_1+1}, Y_{n_1+1}), \ldots, (X_{n_1+n_2}, Y_{n_1+n_2})\}$ where each image has width W and height H
- test input X^{test}
- ML model $f(X, T)$
- parameter values $t_1, \ldots, t_k$
- error rate $0 < \alpha < 1$
- IoU threshold $0 < \tau < 1$
- duplicate prediction IoU threshold $0 < \eta < 1$

Output:

- conformal prediction set with duplicates removed $C_{\hat{\lambda}}(X^{\text{test}})$

// duplicate removal function
procedure UNIQUE*(C, η)
  $m = |C|$
  if $m = 1$ **then return** C
  else
    $C' \leftarrow$ UNIQUE*$(C[1 : m-1], \eta)$
    if there exists $\hat{Y}' \in C'$ such that IoU$(C[m], \hat{Y}') > \eta$ **then return** C'
    else return $C' \cup \{C[m]\}$
    end if
  end if
end procedure

// using n_1 datapoints: compute set S_j of data indices "covered" by each parameter index j (i.e., datapoints where predicted mask $f(X_i, t_j)$ has IoU $> \tau$)
for $j = 1, 2, \ldots, k$ **do**
  for $i = 1, 2, \ldots, n_1$ **do**
    $\rho_{ij} \leftarrow$ IoU$(Y_i, f(X_i, t_j))$
  end for
  $S_j \leftarrow \{i : \rho_{ij} > \tau\}$
end for

// create ordered list L of parameter indices from highest to lowest priority via greedy algorithm
$L \leftarrow []$
for $j' = 1, 2, \ldots, k$ **do**
  // find parameter index that covers greatest number of previously un-covered datapoints
  $L[j'] \leftarrow \arg\max_{j \in [k]} |S_j|$
  for $j'' = 1, 2, \ldots, k$ **do**
    $S_{j''} \leftarrow S_{j''} \setminus S_{L[j']}$
  end for
end for

// using remaining n_2 datapoints: use Conformal Risk Control (CRC) to compute smallest $\hat{\lambda} \in [k]$ such that the first $\hat{\lambda}$ parameter indices in L result in a valid conformal prediction set
$\hat{\lambda} \leftarrow k + 1$
for $\lambda = 1, 2, \ldots, k$ **do**
  for $i = n_1 + 1, \ldots, n_1 + n_2$ **do**
    // compute ordered list of predictions using first λ parameter indices in L and remove duplicates
    $C_\lambda(X_i) \leftarrow \{f(X_i, t_j) : j \in L[1 : \lambda]\}$
    $C_\lambda(X_i) \leftarrow$ UNIQUE*$(C_\lambda(X_i), \eta)$
    // compute loss function for whether $C_\lambda(X_i)$ contains a high IoU prediction
    $l_i(\lambda) \leftarrow \mathbb{1}(\max_{\hat{Y}_i \in C_\lambda(X_i)} \text{IoU}(Y_i, \hat{Y}_i) \leq \tau)$
  end for
  $\hat{r}(\lambda) \leftarrow \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} l_i(\lambda)$ *// empirical risk*
  if $\hat{r}(\lambda) \leq \alpha - \frac{1-\alpha}{n_2}$ **then**
    $\hat{\lambda} \leftarrow \lambda$
    break
  end if
end for

// compute predictions for X^{test} using the first $\hat{\lambda}$ parameter indices in L and remove duplicates
if $\hat{\lambda} = k + 1$ **then**
  $C_{\hat{\lambda}}(X^{\text{test}}) \leftarrow \{0, 1\}^{W \times H}$ *// set of all possible masks*
else
  $C_{\hat{\lambda}}(X^{\text{test}}) \leftarrow \{f(X^{\text{test}}, t_j) : j \in L[1 : \hat{\lambda}]\}$
  $C_{\hat{\lambda}}(X^{\text{test}}) \leftarrow$ UNIQUE*$(C_{\hat{\lambda}}(X^{\text{test}}), \eta)$
end if
return $C_{\hat{\lambda}}(X^{\text{test}})$

We present a modified version of our conformal algorithm that provides finite sample guarantees (Algorithm 3).

We use the notation defined in Section 2 of the main text. For this version of the algorithm, the calibration dataset is randomly split into two sets of size n_1 and n_2: $\{(X_1, Y_1), \ldots, (X_{n_1}, Y_{n_1})\}$ and $\{(X_{n_1+1}, Y_{n_1+1}), \ldots, (X_{n_1+n_2}, Y_{n_1+n_2})\}$. We use the first set of n_1 datapoints to rank the parameter values $t_1, \ldots, t_k$ from highest to lowest priority via a greedy algorithm based on their performance on the data. Then, using the remaining n_2 datapoints, we apply Conformal Risk Control (CRC) (Angelopoulos et al., 2022) to compute the smallest value $\hat{\lambda} \in [k]$ such that the first $\hat{\lambda}$ parameters in the ranked list result in a valid conformal prediction set. The CRC method provides finite sample guarantees for the probability that for a new input X^{test} (from the same distribution as the calibration set), the conformal prediction set $C_{\hat{\lambda}}(X^{\text{test}})$ will contain a mask that has

high IoU with the true mask. We describe the algorithm in more detail below.

Ranking the parameters via greedy algorithm We use the first set of n_1 datapoints as follows. Similar to Algorithm 1, for each t_j, we first compute the set S_j of data indices i where the predicted mask $f(X_i, t_j)$ and true mask Y_i have IoU $> \tau$. We say that S_j is the set of data indices "covered" by the parameter index j.

$$S_j = \big\{i : \text{IoU}\big(Y_i, f(X_i, t_j)\big) > \tau\big\}.$$

Next, we create an ordered list L of parameter indices from highest to lowest priority via a greedy algorithm. At each of k steps, we append to L the parameter index $\arg\max_{j \in [k]} |S_j|$ that covers the greatest number of datapoints that have not already been covered, and then update each of $S_1, \ldots S_k$ to remove the newly covered data indices. This approach is similar to the greedy set cover procedure (Chvatal, 1979; David S Johnson, 1973; Lovász, 1975) used in Line 8 of Algorithm 1 to obtain a set of parameter values that covers $(1 - \alpha)$ fraction of the datapoints. However, in our new algorithm, we do not stop appending parameter indices to L at a specific coverage fraction, but continue until all k parameters have been ranked.

Applying CRC to compute number of parameters to keep Using the remaining n_2 datapoints, we apply CRC to compute the smallest $\hat{\lambda} \in [k]$ such that the first $\hat{\lambda}$ parameters in L result in a valid conformal prediction set. For each $1 \leq \lambda \leq k$, we compute the empirical risk $\hat{r}(\lambda)$ as follows. We loop through the n_2 datapoints; for each point (X_i, Y_i), we compute an ordered list of predictions $C_\lambda(X_i)$ using the first λ parameters from L.

$$C_\lambda(X_i) = \{f(X_i, t_j) : j \in L[1 : \lambda]\}.$$

We remove duplicate predictions using a user-provided IoU threshold $0 < \eta < 1$. The duplicate removal procedure UNIQUE* takes in an ordered list of predictions C and recursively runs duplicate removal on the first $|C| - 1$ elements of C; we call this intermediate output C'. If there exists $\hat{Y}' \in C'$ such that $\hat{Y}'$ and the last element of C have IoU $> \eta$, the procedure returns C'. Otherwise, the procedure returns the union of C' and the last element of C. This ensures that every discarded prediction is within IoU η of one that is kept, and no two predictions that are kept are within IoU η of each other. We have

$$C_\lambda(X_i) \leftarrow \text{UNIQUE*}(C_\lambda(X_i), \eta).$$

We note that unlike the duplicate removal procedure UNIQUE used in Algorithm 2, UNIQUE* ensures that $C_\lambda(X_i) \subseteq C_{\lambda+1}(X_i)$. This will allow us to construct a monotone loss function $l_i(\lambda)$, which is a necessary condition for CRC. For each $\lambda \in [k + 1]$, we define the loss $l_i(\lambda)$ for the datapoint as 1 if none of the predictions in $C_\lambda(X_i)$ have IoU $> \tau$ with the true label Y_i, and 0 otherwise.

$$l_i(\lambda) = \mathbb{1}(\max_{\hat{Y}_i \in C_\lambda(X_i)} \text{IoU}(Y_i, \hat{Y}_i) \leq \tau).$$

For each value of λ, we compute the empirical risk $\hat{r}(\lambda)$ by averaging the losses of the n_2 datapoints. This average is equivalent to the fraction of datapoints where none of the predictions (after duplicate removal) have sufficiently high IoU.

$$\hat{r}(\lambda) = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} l_i(\lambda).$$

If $\hat{r}(\lambda)$ is less than or equal to the CRC threshold of $\alpha - \frac{1-\alpha}{n_2}$, then we set $\hat{\lambda}$ equal to the current value of λ. Otherwise, we continue to iterate through the remaining values of λ until we find a value where the empirical risk is less than or equal to the threshold.

Compute conformal prediction set for new input If there is no value of $\lambda \in [k]$ at which $\hat{r}(\lambda)$ is sufficiently low, then our algorithm returns $\{0, 1\}^{W \times H}$, the set of all possible masks, so that $\hat{r}(k + 1) = 0$. (In practice, to avoid storing $2^{W \times H}$ masks, the algorithm would return an error message.) Otherwise, for a new test input X^{test}, we compute a set of predictions using the first $\hat{\lambda}$ parameters from L.

$$C_{\hat{\lambda}}(X^{\text{test}}) = \{f(X^{\text{test}}, t_j) : j \in L[1 : \hat{\lambda}]\}.$$

We remove duplicates to obtain the final conformal prediction set

$$C_{\hat{\lambda}}(X^{\text{test}}) \leftarrow \text{UNIQUE*}(C_{\hat{\lambda}}(X^{\text{test}}), \eta).$$

The resulting finite-sample guarantee is that, for a new test input X^{test}, with probability at least $1 - \alpha$, there exists some $\hat{Y}^{\text{test}} \in C_{\hat{\lambda}}(X^{\text{test}})$ with $\text{IoU}(Y^{\text{test}}, \hat{Y}^{\text{test}}) > \tau$. This is stated formally in the following theorem.

Theorem H.1. *Let $C_{\hat{\lambda}}(X^{\text{test}})$ denote the set of segmentation masks returned when running Algorithm 3 with calibration samples $\big((X_i, Y_i)\big)_{i=1}^{n_1}$ and $\big((X_i, Y_i)\big)_{i=n_1+1}^{n_1+n_2}$ and parameters $\alpha, \tau, \eta \in (0, 1)$. Under Assumption 1 (calibration and test samples are i.i.d.),*

$$\mathbb{P}\Big(\max_{\hat{y} \in C_{\hat{\lambda}}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}}, \hat{y})\big\} > \tau\Big) \geq 1 - \alpha.$$

Proof. For each calibration point (X_i, Y_i), the loss function $l_i(\lambda) = \mathbb{1}(\max_{\hat{Y}_i \in C_\lambda(X_i)} \text{IoU}(Y_i, \hat{Y}_i) \leq \tau)$ is non-increasing in λ. This is because by design $C_\lambda(X_i) \subseteq C_{\lambda+1}(X_i)$, so the maximum IoU of the prediction set can only improve as λ increases. Furthermore, we define $C_{k+1}(X_i)$ as the set of all possible masks $\{0, 1\}^{W \times H}$, which is guaranteed to contain the true mask, so $l_i(\lambda_{\text{max}}) = l_i(k + 1) = 0 < \alpha$. Thus, we can apply Conformal Risk Control (Angelopoulos et al., 2022).

Our algorithm computes the smallest value of λ such that the empirical risk $\hat{r}(\lambda) = \frac{1}{n_2} \sum_{i=n_1+1}^{n_1+n_2} l_i(\lambda)$ falls below the CRC threshold:

$$\hat{\lambda} = \min\{\lambda : \hat{r}(\lambda) \leq \alpha - \frac{1 - \alpha}{n_2}\}.$$

For a new test point $(X^{\text{test}}, Y^{\text{test}})$, let $l_{\text{test}}(\lambda) = \mathbb{1}(\max_{\hat{y} \in C_\lambda(X^{\text{test}})} \text{IoU}(Y^{\text{test}}, \hat{y}) \leq \tau)$. Then by Theorem 1 of Angelopoulos et al. (2022),

$$\mathbb{P}\Big(\max_{\hat{y} \in C_{\hat{\lambda}}(X^{\text{test}})} \big\{\text{IoU}(Y^{\text{test}}, \hat{y})\big\} \leq \tau\Big) = \mathbb{E}[l_{\text{test}}(\hat{\lambda})] \leq \alpha.$$

□

We note that the parameter-ranking step using the first n_1 datapoints is not necessary for our algorithm to be valid, since CRC can be applied regardless of how L is ordered. However, if we do not rank the parameters based on their performance on the data, $\hat{\lambda}$ may be very large, i.e. we may need to keep many parameters in order to ensure sufficiently low empirical risk; this is computationally costly and may result in large prediction sets that are inconvenient to use.

Comparison to asymptotic-guarantee algorithm The finite-sample-guarantee version of the algorithm will generally give similar results as the asymptotic-guarantee version presented in Section 2 of the main text. This is because the greedy algorithm used in the parameter-ranking step of Algorithm 3 is closely related to the greedy set cover algorithm used in Line 8 of Algorithm 1. In Algorithm 1, we use the greedy set cover algorithm to find a set of parameters that covers $(1 - \alpha)$ fraction of the calibration points; analogously, in the CRC step of Algorithm 3, we add parameters from the ranked list L until we cover $(1 - \alpha + \frac{1-\alpha}{n_2})$ fraction of the n_2 calibration points.

However, one advantage of Algorithm 1 is that we can improve the greedy set cover by brute-forcing over all smaller subsets of $[k]$ in order of cardinality until we either find a true minimal-size set cover or confirm that the greedy set cover is already minimal. This is not possible in Algorithm 3 because we are restricted to parameter sets of the form $L[1 : \lambda]$, i.e., adding parameters in order of their greedy algorithm ranking. Thus, Algorithm 3 cannot guarantee that the set of parameters used to produce the conformal prediction set is minimal-size.

Another advantage of the asymptotic algorithm is that the duplicate removal procedure $\text{UNIQUE}(C, \eta)$ used in Algorithm 2 brute-forces over all subsets of C in order of increasing cardinality until it finds a *minimal-size* subset such that every discarded prediction is within IoU η of one that is kept. In contrast, the duplicate removal procedure $\text{UNIQUE}^*(C, \eta)$ used in Algorithm 3 outputs a subset of predictions such that every discarded prediction is within IoU η of one that is kept and no two predictions that are kept are within IoU η of each other, but this subset is not necessarily minimal-size. ($\text{UNIQUE}^*(C, \eta)$ ensures that even after duplicate removal, $C_\lambda(X_i) \subseteq C_{\lambda+1}(X_i)$ and thus the maximum IoU-based loss function $l_i(\lambda)$ is non-increasing in λ, which is a necessary condition for CRC.) As a result, the final prediction sets produced by Algorithm 3 may be larger than those from the asymptotic algorithm.

We note that the finite-sample algorithm has the advantage of computational efficiency because its greedy parameter ranking procedure and duplicate removal procedure run in polynomial time, making it useful in cases where brute-forcing is not computationally tractable.

I DATASETS AND MODELS

Agricultural field delineation Delineating agricultural fields from satellite imagery is an important task for monitoring land use and supporting precision agriculture. The state-of-the-art uses deep semantic segmentation models, such as U-Nets and their variants, followed by post-processing steps like watershed segmentation or connected components analysis to separate adjacent fields (Kerner et al., 2025; Waldner et al., 2021).

Each satellite image contains multiple field instances. We use 350 images and instance segmentation labels from the Fields of the World (FTW) test dataset for France (Kerner et al., 2025), and we randomly divide the data into 250 calibration images and 100 test images. We randomly sample 50 pixels from each image and only keep pixels that are in a field, resulting in 2476 calibration pixels and 917 test pixels.

For predictions, we use the pretrained FTW U-Net model provided by Kerner et al. (2025) to estimate field boundary probabilities at each pixel, followed by a watershed algorithm to convert the probability maps into field instances. The tunable parameter T corresponds to the watershed threshold: as T increases, neighboring fields are merged, yielding larger predicted instances. We consider T values from 0 to 0.009 in steps of 0.001, and from 0.01 to 1 in steps of 0.01.

Cell segmentation Segmenting individual cells from microscopy images is a common task in biology and medicine, supporting cell counting and morphology characterization. A widely used approach is Cellpose (Stringer et al., 2021), which has recently been extended with Segment Anything (Cellpose-SAM) (Pachitariu et al., 2025) to improve robustness across diverse imaging conditions.

We evaluate our method on 52 grayscale images with instance segmentation labels from the Cellpose test dataset. We split the data into 30 calibration images and 22 test images and randomly sample 50 pixels from each image, only keeping pixels in a cell (or other object) instance. This results in 925 calibration pixels and 659 test pixels.

For predictions, we use the Cellpose-SAM model to segment each image into cell instances. The tunable parameter T corresponds to the "cell extent" threshold: increasing T causes more pixels to be classified as background, producing smaller predicted instances. We sweep T from –5 to 5 in steps of 1.

Vehicle detection Detecting vehicles in street-level imagery is a central task in autonomous driving and traffic monitoring.

We evaluate our method on 200 images with car instance labels from the Cityscapes street scene validation dataset (Cordts et al., 2016). We split the data into 100 calibration and 100 test images and randomly sample 20 pixels from each image, only keeping pixels in the *car* class. This yields 105 calibration and 121 test pixels.

For predictions, we use the Segment Anything Model (SAM) (Kirillov et al., 2023), which takes an RGB image and a pixel coordinate and returns mask probability maps for the object containing that pixel. SAM produces three candidate mask probability maps per query, ranked by confidence. There are two tunable parameters: T_1, which indexes the candidate mask $(1, 2, 3)$, and T_2, the mask probability threshold. As T_2 increases, more pixels are classified as not being in the mask, so the predicted object instances become smaller. We iterate over T_1 values in $\{1, 2, 3\}$ and T_2 values from 0 to 1 at increments of 0.05.

J EXPERIMENTAL VALIDATION OF FINITE SAMPLE VERSION OF ALGORITHM

We perform an empirical comparison of the asymptotic and finite sample guarantee versions of our conformal algorithm. We run the finite sample algorithm (Algorithm 3) using the same error rate α and IoU threshold τ as in the original experiments from the main text, as well as the same calibration and test sets. The finite sample algorithm requires randomly splitting the calibration pixels into two sets; we use a 50-50 split.

We evaluate the coverage of the finite sample algorithm on the test set and compare it to the coverage of the asymptotic version (Table 3). Similar to the asymptotic version, the finite sample coverages are close to the target coverages $(1 - \alpha)$, as expected. We note that the finite sample coverages are computed using the IoU threshold τ instead of $\tilde{\theta}$ (which is used for the asymptotic version) since the finite sample algorithm does not need to re-calibrate the IoU threshold after duplicate removal.

Table 3: **Asymptotic and Finite Sample Algorithm Coverages of Test Set Predictions.**

EXAMPLE	τ	$\tilde{\theta}$	TARGET COVERAGE $(1 - \alpha)$	ASYMPTOTIC COVERAGE	FINITE-SAMPLE COVERAGE
Field delineation	0.7	0.696	0.8	0.797	0.779
Cell segmentation	0.75	0.760	0.8	0.835	0.816
Vehicle detection	0.8	0.842	0.9	0.843	0.893

Furthermore, we visualize the distribution of conformal prediction set sizes for test data points after removing duplicate predictions (Figure J.1). Like the asymptotic algorithm, the finite sample algorithm results in diverse set sizes after duplicate removal, adapting to query difficulty (although the maximum prediction set sizes differ slightly for the asymptotic and finite sample algorithms in the field delineation and cell segmentation examples).

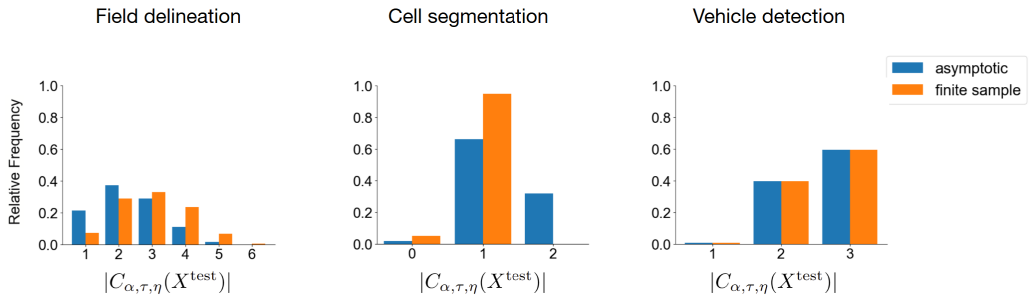

Figure J.1: **Distribution of conformal prediction set sizes on test set for asymptotic and finite sample versions of the algorithm.**

K SENSITIVITY ANALYSES ON CONFORMAL ALGORITHM PARAMETERS

We run experiments to analyze the robustness of our algorithm, including sensitivity analyses on error rate α, IoU threshold τ, duplicate removal IoU threshold η, tunable parameter space size k, and calibration set size n (Figure K.1).

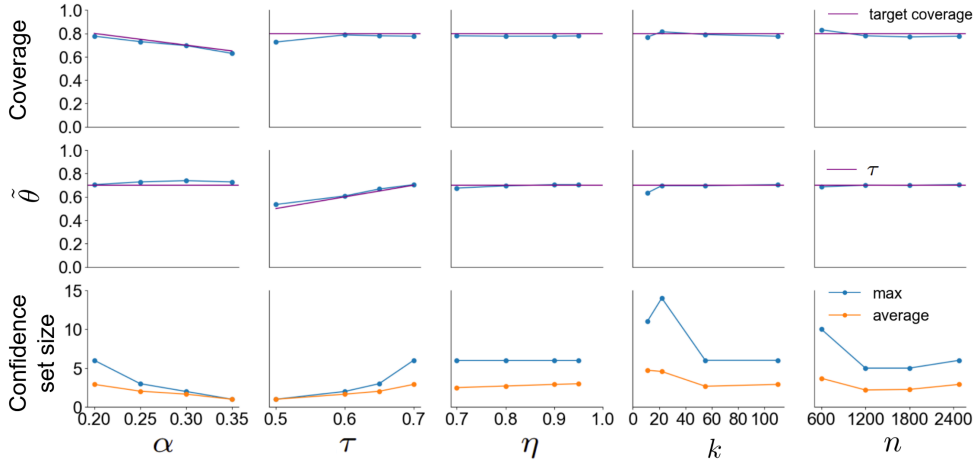

Figure K.1: **Sensitivity analyses on conformal algorithm parameters: error rate α, IoU threshold τ, duplicate removal IoU threshold η, tunable parameter space size k, and calibration set size n.**

We use the field delineation dataset with the same default parameters as in the main text: $\alpha = 0.2$, $\tau = 0.7$, $\eta = 0.9$, $k = 110$, and $n = 2476$. For each experiment, we vary one parameter while holding the others constant. We use the calibration set to compute a confidence set for the tunable parameter, re-calibrate the IoU threshold after duplicate removal from τ to $\tilde{\theta}$, and evaluate performance on the test set of 917 points. For each experiment, we record $\tilde{\theta}$, the test coverage, maximum prediction set size, and average prediction set size on the test set. (For all experiments, we use the asymptotic-guarantee version of our algorithm with the greedy set cover subprocedure for computational efficiency.)

Varying error rate α We vary error rate $\alpha = 0.2, 0.25, 0.3, 0.35$. For all experiments, coverage is close to the target coverage of $1 - \alpha$, showing that our method is robust to changes in α. The re-calibrated IoU threshold $\tilde{\theta}$ stays close to $\tau = 0.7$ as expected. As α increases, the prediction set size decreases because fewer predictions are required in order to attain the desired $1 - \alpha$ coverage.

Varying IoU threshold τ We vary IoU threshold $\tau = 0.5, 0.6, 0.65, 0.7$. For $\tau = 0.5$, the coverage (0.73) is slightly lower than the target coverage of $1 - \alpha = 0.8$; we believe this is due to random variation between the calibration and test sets. For all other experiments, coverage is close to the target coverage. The re-calibrated IoU threshold $\tilde{\theta}$ stays close to τ as expected, for all values of τ. As τ increases, the prediction set size increases because more predictions are required in order to attain IoU greater than τ on at least one prediction in the set.

Varying duplicate removal IoU threshold η We vary duplicate removal IoU threshold $\eta = 0.7, 0.8, 0.9, 0.95$. For all experiments, coverage is close to the target coverage of $1 - \alpha = 0.8$, showing that our method is robust to changes in η. The re-calibrated IoU threshold $\tilde{\theta}$ stays close to τ; even for η as low as 0.7, $\tilde{\theta}$ remains fairly high (0.675), so duplicate removal only has a minor effect on whether the final prediction set contains a high enough IoU prediction. The maximum prediction set size stays the same because it is calibrated using α and τ only and does not depend on η (see Algorithm 1). However, the average prediction set size increases from 2.5 to 3.0 because fewer elements are removed during duplicate removal.

Varying tunable parameter space size k We vary tunable parameter space size $k = 11, 22, 55, 110$. (In the paper, we use $k = 110$ tunable parameter values from 0 to 0.009 in steps of 0.001, and from 0.01 to 1 in steps of 0.01. For the sensitivity analysis, we create sets of $k = 11, 22, 55$ tunable parameter values by increasing the step sizes by a factor of 10, 5, and 2 respectively.) For $k = 11$, no subset of the parameters could attain the target IoU threshold $\tau = 0.7$ on $1 - \alpha = 0.8$ fraction of calibration points, so we used the full set of 11 parameters in lieu of a confidence set. This resulted in a lower re-calibrated IoU threshold ($\tilde{\theta} = 0.63$). The re-calibrated IoU threshold $\tilde{\theta}$ stays close to $\tau = 0.7$ for all other values of k. For all experiments, coverage at $\tilde{\theta}$ is close to the target coverage of $1 - \alpha = 0.8$. Finally, when k is small, the maximum and

average prediction set sizes are large because with a less granular tunable parameter space, more parameters are needed in order to attain high maximum IoU (and when $k = 11$, we are unable to attain IoU > 0.7 on 80% of the calibration points even if we use all the parameters).

Varying calibration set size n We vary calibration set size $n = 600, 1200, 1800, 2476$. For all experiments, coverage is close to the target coverage of $1 - \alpha = 0.8$. The re-calibrated IoU threshold $\tilde{\theta}$ stays close to $\tau = 0.7$ as expected. However, for the smallest value $n = 600$, the maximum and average prediction set sizes are substantially larger than for the other values of n, indicating that the calibration results are less stable at small calibration set sizes; the computed confidence set of tunable parameter values is more sensitive to random variation in the calibration data when n is small. When we tried further decreasing n to 300, no subset of the parameters could attain IoU > 0.7 on 80% of calibration points.