Learning to Count Mosquitoes for the Sterile Insect Technique

Yaniv Ovadia
Google Inc.
yovadia@google.com

Yoni Halpern
Google Inc.
yhalpern@google.com

Dilip Krishnan
Google Inc.
dilipkay@google.com

Josh Livni
Verily Inc.
jlivni@google.com

Daniel Newburger
Verily Inc.
dnewb@google.com

Ryan Poplin
Verily Inc.
rpoplin@google.com

Tiantian Zha
Verily Inc.
zha@google.com

D. Sculley
Google Inc.
dsculley@google.com

ABSTRACT

Mosquito-borne illnesses such as dengue, chikungunya, and Zika are major global health problems, which are not yet addressable with vaccines and must be countered by reducing mosquito populations. The Sterile Insect Technique (SIT) is a promising alternative to pesticides; however, effective SIT relies on minimal releases of female insects. This paper describes a multi-objective convolutional neural net to significantly streamline the process of counting male and female mosquitoes released from a SIT factory and provides a statistical basis for verifying strict contamination rate limits from these counts despite measurement noise. These results are a promising indication that such methods may dramatically reduce the cost of effective SIT methods in practice.

CCS CONCEPTS

•Applied computing → Mathematics and statistics; Imaging;

KEYWORDS

image modeling, quality assurance, counting from images

1 INTRODUCTION

Mosquitoes kill over a million people per year and sicken hundreds of millions more [22, 23]. *Aedes aegypti* alone is the primary vector for dengue and can additionally transmit Zika and chikungunya. Currently, their populations are usually controlled via pesticides and tedious elimination of standing water. The Sterile Insect Technique (SIT) is a promising approach in which an overwhelming number of sterile insects are released to compete with the wild population for mates and ultimately reduce the rate of new offspring [2, 4, 7, 8, 12].

Only female mosquitoes bite and transmit disease, so it is important for SIT implementations to minimize the number of released females and to accurately measure the female contamination rates of their releases. Therefore robust procedures for correctly classifying mosquito sex are paramount to the success of this technique.

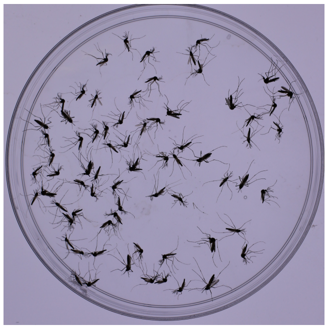

Figure 1: A petri dish of male and female immobilized mosquitoes from which we expect our model to count the total number of insects and localize females.

Existing methods for sex sorting in mosquito SIT trials include mechanical separation by pupal size, insecticide laced bloodmeals, genetic sexing, and manual separation. Quality assurance is generally a manual process [4, 7, 8, 19].

This paper presents a quality assurance system for measuring the female contamination rate of an *Aedes aegypti* SIT factory using a machine-learned image model to assist a human technician with counting immobilized male and female mosquitoes in petri dishes.

After detailing the model and its training data, we report results under various hyperparameter settings and describe the metrics we use to quantify a model's quality. In Section 7 we analyze our model's quality relative to theoretical limits and interpret them with respect to practical considerations. Specifically, we investigate the limitations of patch-based detection, derive a statistical basis for performing validation, and derive bounds on statistical power with which we evaluate the effect of imperfect detection on minimum sample sizes.

2 QUALITY ASSURANCE PROCESS

The proposed QA procedure entails imaging petri dishes of immobilized mosquitoes and counting from these images. An example image is depicted in Figure 1. Our model estimates mosquito counts and detects females on small image patches. It aggregates these

KDD'17, August 13–17, 2017, Halifax, NS, Canada.
© 2017 Copyright held by the owner/author(s). 978-1-4503-4887-4/17/08.
DOI: http://dx.doi.org/10.1145/3097983.3098204

counts into a total count and provides approximate female localizations. A human technician is then responsible for counting any detected females. By only leaving female counting to the technician, we minimize the risk of miscounting females while expending minimal human effort.

These tallies enable one to infer the factory's contamination rate. Since technicians are not expected to inspect non-flagged petri dish regions, the detection model's sensitivity is crucial for generating accurate female counts.

3 BACKGROUND

As summarized in [14], contemporary approaches to counting from images usually employ counting by detection or regression.

Detection-based solutions attempt to localize all instances of the relevant object and yield the number of such detections. This usually entails predicting a confidence map over the image and applying a combination of thresholding and non-maximum suppression to isolate individual items. However, other detection approaches, such as semantic segmentation of distinct items, can be similarly employed [9, 17]. While these methods offer more interpretability, they can struggle with images with overlapping items as is the case in our petri dishes of mosquitoes.

Counting by regression skips localization and simply learns a mapping from the image to the number of visible objects. These methods vary in training annotations (e.g. bounding boxes or point localizations from one or more raters) and regression targets, which may simply be the integer count of wholly contained items as in [18], or the sum over a density map. These density maps represent fractional counts over a region of the image. For example, [5] used bounding box annotations and defined a density map with counts uniformly distributed over those boxes, while [14] and [24] employed a sum of Gaussian kernels. On a dataset with noisy point localizations, [3] defined a density map that distributed each item's count over a region defined by multiple noisy point annotations.

Most modern approaches to counting from images use deep convolutional neural network architectures [3, 5, 9, 24]. Deep convolutional neural networks have shown tremendous success in a number of computer vision tasks, including image classification [13, 21], video classification [10], object-detection [20], and semantic segmentation [15]. In a convolutional neural network, successive layers of small tunable filters are trained to extract relevant features from an image. At each layer, a number of filters are applied in a convolutional manner across the spatial dimensions of the image and their outputs are pooled and nonlinearly transformed before being passed to the next layer. In contrast to machine learning systems that use carefully engineered image descriptors such as HOG [6] and SIFT [16], convolutional neural networks learn relevant features as part of the supervised learning task.

Since overlapping insects are common in our dataset, we opted to count by regression. Specifically, our model predicts the sum of (possibly fractional) line-segments present in each patch that makes up the complete image. We also supply the CNN with additional surrounding pixels for context and an indicator mask to clarify the boundary between the context region and region of interest. A similar context region (without an indicator mask) was used to count cars from overhead imagery in [18].

4 TRAINING DATA

Our dataset consists of 500, 8000 × 8000 pixel petri dish images containing approximately 200 mosquitoes each at a male:female ratio of approximately 10:1. This ratio is much lower than will be the case in practice, but the more balanced class ratio is helpful to provide sufficient information value during training.

We collect ground-truth annotations for training using two phases of human labeling. In the first phase, raters localize mosquitoes with two endpoints — one at the head and another at the end of the abdomen. This localization is verified by another rater before proceeding to phase two (else it returns to the first phase for correction). In phase two, we ask multiple raters to classify each localized mosquito as male, female, or unknown from image crops centered around the localization.

Female mosquitoes are distinguished by larger bodies and little-to-no feathering on their antennae; these features are easily visible in Figure 2. In cases where raters could not determine a mosquito's sex, we err on the side of caution and treat those instances as females.

5 MODEL

Since we wish to produce counts and approximate localizations, we employed a patch-based CNN to make predictions on small square patches with dimensions $R \times R$. To generate fractional count labels for each patch, we linearly interpolate M points between each annotated segment and sum the number of these points contained in the focus region. The detection label is positive if any of these points corresponding to a female mosquito is contained in the focus region. We also support assigning distinct interpolation sizes for counting (M_{ctr}) and detection (M_{det}).

To avoid confusing the model with images of small fractions of mosquitoes that are insufficient for sex determination, we pad the input image with a *context region* of C additional pixels surrounding the $R \times R$ *focus region*; the total input with dimensions $S \times S$ (where $S = R + 2C$) will be referred to as the *model window*. We convey the boundaries between the focus and context regions to the model by concatenating a context vs. focus indicator-mask to the window's color channels.

Our network is an instance of the InceptionV3 architecture [21] trained from scratch with randomly initialized parameters and filter bank sizes scaled by 0.3, yielding a shared hidden layer from which we make two predictions: (1) *Female detection*: true if any portion of a female mosquito is present in the focus region. (2) *Counting*: the sum of fractional mosquitoes contained in the focus region. To train the model, we minimize the sum of L_2 losses from counting predictions [1] and logistic loss from female detection. We observed the two loss values to be relatively comparable (within one or two orders of magnitude), so neither objective overwhelmingly dominates shared parameter optimization.

At inference, we run the model on $S \times S$-sized patches with stride R. Counts are summed across all patches, and detection predictions exceeding a configured threshold alert a human operator who will inspect the patch and update the females count if appropriate.

[1] Subsequent experiments suggest that count predictions from softmax classification similar to that employed by [18] are more accurate than counts from minimizing L_2 loss.

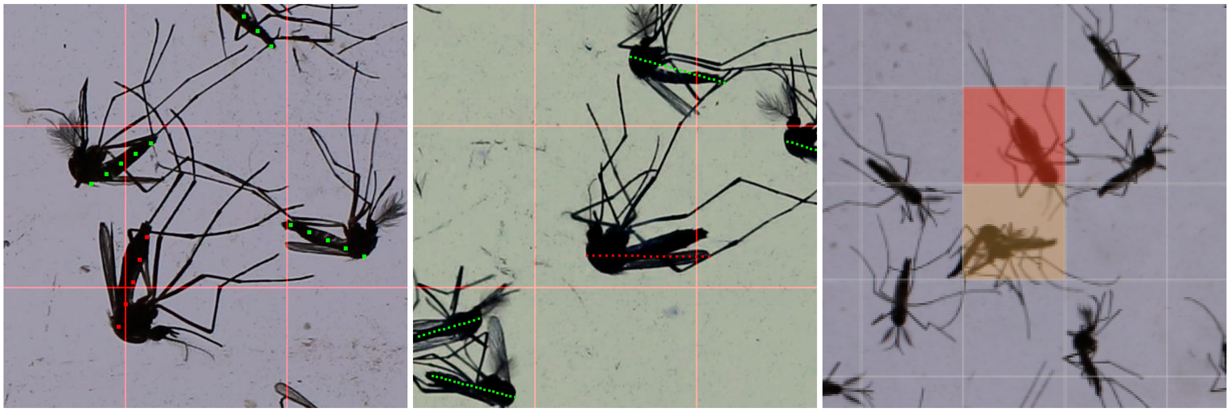

Figure 2: Left: Two input examples with gridlines surrounding the focus-region and red and green interpolated points denoting ground-truth female and male localizations. Right: An example detection.

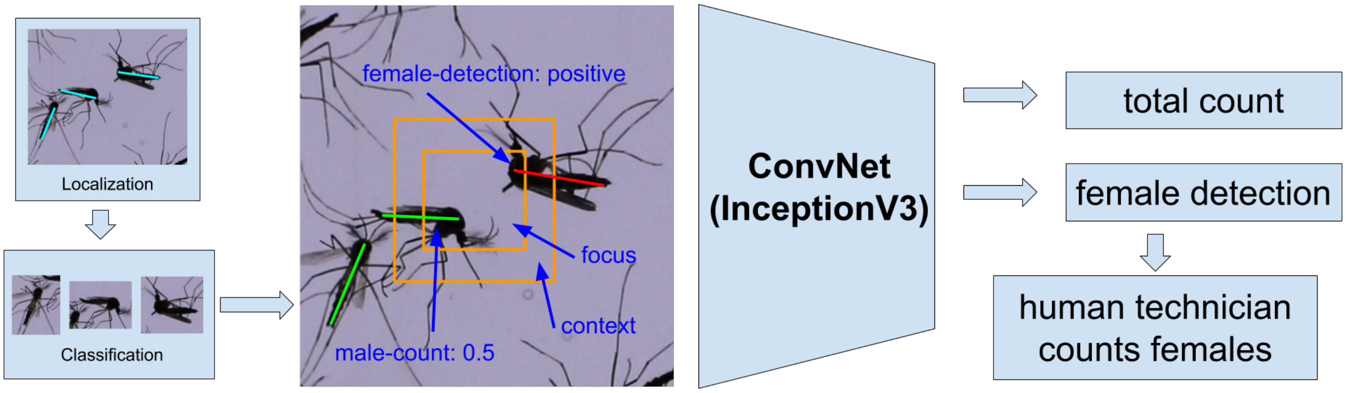

Figure 3: A flow chart depiction of the model and its training data annotation process. Human labelers begin by localizing each mosquito with two endpoints, and each localized insect is then cropped and sent to multiple raters for male/female classification. Random, dynamically extracted image patches are generated at training and labeled according to the contents of the focus region. At serving, the female detections assist human technicians by localizing female insects to count.

5.1 Training

We trained the model using distributed TensorFlow with AdamOptimizer for gradient descent [1, 11]. Each training worker dynamically extracted random patches and employed a shuffling queue to avoid correlations within minibatches. Per conventional practice with deep CNNs, we applied dropout and random image distortions including random flips and random adjustments to brightness, contrast, hue and saturation.

6 RESULTS

Simple quality metrics of per-patch prediction accuracy do not easily translate to actionable interpretations and are not comparable between some hyperparameter settings. We begin by describing the quality metrics we will use to compare models and proceed to report results from experiments studying the effects of the context mask as well as focus-size and the number of interpolation points used to represent insects for patch labeling.

Models in these experiments used a fixed context size of $C = 150$ as this roughly corresponds to the length of a mosquito in pixels. Other experimental parameters varied between the studies and are noted in Appendix A.

6.1 Model Quality Metrics

To quantify counting quality, per-plate counting accuracy will suffice, but translating patch detections into plate-level counting metrics is less obvious.

We propose scoring each mosquito according to the maximum detection probability assigned to patches containing any part of its corresponding segment and evaluate those scores according to specificity-at-fixed-sensitivity. This metric is intended to capture the performance of female counting when coupled with a human technician since we expect the technician to inspect each insect touching a positive patch.

One implication of this metric is that specificity is bounded by the resolution enabled by the model's focus size (R) as patches that touch both male and female mosquitoes can confound perfect scoring. We investigate this effect in detail in Section 7.1.

6.2 Context Region

As described in Section 5, we include C pixels of padding around the focus region to enable the model to distinguish between male and female mosquitoes without identifying features (i.e. antennae) contained in the focus region. Additionally, we concatenate to the input image patches a fourth color channel with $\{0, 1\}$ indicators denoting which pixels belong to focus vs context.

We studied the effect of this complexity by training three otherwise identical models: one without a context region, one with a context region, and one using both a context region and context mask channel.

Models in these experiments fixed focus size at $R = 200$, and interpolation sizes at $M_{\text{ctr}} = 10$ and $M_{\text{det}} = 1$. Each model trained for 4 million steps

The results, summarized the table below, suggest that the context buffer is useful for both counting and detection, but the context mask offers minimal improvement at best.

model	specificity @ 95% sensitivity	RMSE
no context	93.61%	22
with context	94.67%	14
with context & mask	94.60%	12

6.3 Focus size and interpolation size

To tune model quality, we ran a hyperparameter grid search over R (focus size) and M (interpolation size).

The tables below lists per-mosquito specificities at 95% sensitivity for the female detector:

	$R = 200$	$R = 300$
$M = 1$	91.3%	87.1%
$M = 2$	89.7%	83.7%
$M = 10$	90.1%	82.9%

and root-mean-squared-error for the counting model:

	$R = 200$	$R = 300$
$M = 1$	15.4	n/a
$M = 2$	6.1	5.0
$M = 10$	2.9	1.7

The $R = 300$, $M = 1$ case uniformly predicted zero counts, which presumably indicates that final hidden layer activations became stuck in the flat domain of the ReLU output activation function. For study details, see Appendix A.2.

We observe from these experiments that the larger focus size ($R = 300$) improves counting accuracy but harms detection sensitivity. Counting accuracy also favors larger interpolation sizes; while the pattern for detection is less obvious, single point representations perform best.

7 ANALYSIS

In this section, we analyze our model quality results relative to theoretical limits and interpret them with respect to practical considerations. We begin by comparing the female-detector's sensitivity to theoretical limits imposed by using a model on fixed size patches.

Next, we proceed to analyze the implications of our model's quality on its viability as a statistically sound tool for quality assurance. First, we introduce notation and the underlying dynamics we assume to govern factory contamination and our model's observations. In Section 7.3, we describe how to use our system's noisy counting measurements to prove that a factory is operating below a maximum contamination rate. From this validation policy and the dynamics model, we can derive bounds for our system's type I error probability, from which we can bound our system's sampling overhead compared to the sampling requirements of a hypothetical counting system with perfect accuracy.

7.1 Limitations of patch-based detection

The per-insect specificity-at-sensitivity metric described in Section 6.1 is naturally affected by focus-size R and by the number of points we use to represent each mosquito's line segment M. Larger focus-sizes and larger M imply more overlap between mosquitoes corresponding to each detection prediction. For example, a female insect for which each relevant patch is also occupied by a male mosquito is impossible to classify without some loss to sensitivity or specificity. Furthermore, by framing our model's learning task to predict positive detection when any part of a female is present in a patch, we further limit the its ability to optimize this metric directly.

To study this effect, we evaluated specificity of patch scores directly derived from ground-truth annotations where patches containing a female mosquito point are positive and patches without a point corresponding to a female insect are negative. These results are listed in the table below.

	$R = 100$	$R = 200$	$R = 300$
$M = 1$	99.8%	97.6%	93.0%
$M = 2$	98.0%	91.4%	84.1%
$M = 3$	97.6%	90.8%	83.6%
$M = 10$	97.1%	90.5%	83.2%

We also evaluated a more fine-grained patch scoring function that assigns a score proportional to the number of female insect segment points observed in each patch. The table below lists measured specificities at 90% sensitivity.

	$R = 100$	$R = 200$	$R = 300$
$M = 1$	99.9%	97.7%	94.3%
$M = 2$	98.1%	93.1%	87.4%
$M = 3$	97.4%	92.2%	91.5%
$M = 10$	98.9%	95.6%	91.9%

As expected, larger patches have lower resolution with which to identify female insects without flagging false positives. Without fine-grained scoring, fewer points also naturally improves resolution as at most one positive patch is assigned per female, but when we allow non-binary scores, additional points enable one to select a better threshold.

These specificities represent upper bounds on our detection model's quality, but since the male:female ratio in our dataset is much lower than we expect to see in practice, we can expect the specificity loss that results from limited resolution to reduce in practice.

7.2 Modeling the QA process

Our validation and statistical power calculations assume the following underlying dynamics model.

We begin by listing some notation.

- k is the number of measured petri dishes with an average of m mosquitoes per dish
- n and n_{obs} are the true and measured number of sampled mosquitoes respectively
- f and f_{obs} are the true and measured number of female mosquitoes
- σ_d^2 is the per-plate mosquito count variance
- σ_m^2 is the counting model's mean squared error
- s is the detection model's sensitivity
- r is the factory's contamination rate

The following process describes our assumptions of the system's underlying dynamics

1. $n \sim \mathcal{N}(km, k\sigma_d^2)$ — Each plate consists of a normally distributed number of mosquitoes, so the total number of insects is also normally distributed with variance proportional to the number of petri dishes.
2. $n_{\text{obs}} \sim \mathcal{N}(n, k\sigma_m^2)$ — We assume the counting model's measurement error is Gaussian with variance σ_m^2 per plate.
3. $f \sim \text{Binomial}(n, r)$ — Each sampled mosquito is female with probability determined by the factory's contamination rate.
4. $f_{\text{obs}} \sim \text{Binomial}(f, s)$ — Each female is observed with probability corresponding to our model's per-mosquito sensitivity.

For simplicity, we exclude the model's counting predictions from our analysis, so we assume that total counts are derived only from the number of sampled plates. In practice, we can take advantage of the model's more accurate count measurements to yield a more efficient procedure.

7.3 Validating a factory from measurements

In this section we derive conditions for validating a factory as having contamination rate below some maximum r^* such that the probability of misclassifying a non-compliant factory (type II error) is at most ϵ_{II}.

Given that we confirmed a count of f_{obs} females from k petri dishes, we will compare f_{obs} to a threshold female count f^* such that the probability of observing fewer than f^* female mosquitoes from a factory with an excessive contamination rate is less than ϵ_{II}.

$$P_{\text{err II}} = P(f_{\text{obs}} < f^* | r > r^*) \le \epsilon_{\text{II}}$$

First, observe that type II errors are most probable at minimal r, so $P(f_{\text{obs}} < f^* | r > r^*) \le P(f_{\text{obs}} < f^* | r = r^*)$. We then marginalize over the true sample count, $n \sim N(km, k\sigma_d^2)$, splitting the domain of integration into n below and above n^{lo}, and bounding the probability of committing a type II error in the lower domain by 1.

$$\begin{aligned} P_{\text{err II}} &\le \mathbf{E}_{n\sim\mathcal{N}(km,k\sigma_d^2)}\left[P(f_{\text{obs}} < f^* | r = r^*)\right] \\ &\le \int_{n \le n^{\text{lo}}} P(n)(1) + \int_{n > n^{\text{lo}}} P(n)P(f_{\text{obs}} < f^* | r = r^*) \end{aligned}$$

Note that, $f_{\text{obs}} \sim \text{Binom}(n, r^*s)$, but we can shift this distribution to n^{lo} to upper bound $P(f_{\text{obs}} < f^*)$. It follows that

$$P_{\text{err II}} \le 1 - q_{\text{II}} + q_{\text{II}} P(f_{\text{obs}} \sim \text{Binom}(n^{\text{lo}}, r^*s) < f^*)$$

where we define $q_{\text{II}} = P(n > n^{\text{lo}})$.

It follows that assigning f^* such that

$$\sum_{f_{\text{obs}}=0}^{f^*-1} \text{Binom}(f_{\text{obs}} | n^{\text{lo}}, r^*s) \le \frac{\epsilon_{\text{II}} + q_{\text{II}} - 1}{q_{\text{II}}}$$

will ensure that the probability of validating a non-compliant factory is less than ϵ_{II}.

7.4 Statistical power

To evaluate our model's viability as a QA solution, we measure the loss to statistical power that results from its imperfect counts; specifically, we will derive a bound on the probability of invalidating a compliant factory (type I error) and use this to compute a minimal sample size of petri dishes k that satisfies

$$P_{\text{err I}} = P(f_{\text{obs}} \ge f^* | r < r^*) < \epsilon_{\text{I}}$$

As in Section 7.3, we begin by marginalizing over $n \sim \mathcal{N}(km, k\sigma_d^2)$ and splitting the domain of integration.

$$\begin{aligned} P_{\text{err I}} &= \mathbf{E}_n\left[P(f_{\text{obs}} \ge f^* | r < r^*)\right] \\ &\le 1 - P(n < n^{\text{hi}}) + \int_{n < n^{\text{hi}}} P(n)P(f_{\text{obs}} \ge f^* | r < r^*) \end{aligned}$$

As earlier, we define $q_{\text{I}} = P(n < n^{\text{hi}})$ and shift $f_{\text{obs}} \sim \text{Binom}(n, rs)$ to $\text{Binom}(n^{\text{hi}}, rs)$.

$$P_{\text{err I}} \le 1 - q_{\text{I}} + q_{\text{I}} P(f_{\text{obs}} \sim \text{Binom}(n^{\text{hi}}, rs) \ge f^* | r < r^*)$$

Thus we can bound the probability of invalidating a compliant factory at ϵ_{I} by selecting a sufficiently large sample size, k, that

$$\frac{1 - \epsilon_{\text{I}}}{q_{\text{I}}} \le \sum_{f_{\text{obs}}=0}^{f^*} \text{Binom}(f_{\text{obs}} | n^{\text{hi}}, rs)$$

7.5 Solving for a minimum sample size

A bound on the type I error probability has limited interpretability and practical utility on its own. Instead, we wish to reason about the real-world viability of our model in terms of the sample size required to perform validations subject to fixed tolerances on type I and II error probabilities as this sampling overhead can easily translate to monetary or operational efficiency costs.

Formally, we wish to compute

$$k^*(r) = \min\{k | \forall k' \ge k, \exists f^*, \tilde{P}_{\text{err II}}(f^*, k') \le \epsilon_{\text{II}}, \tilde{P}_{\text{err I}}(f^*, k') \le \epsilon_{\text{I}}\}$$

where $\tilde{P}_{\text{err II}}$ and $\tilde{P}_{\text{err I}}$ are the upper bounds on type I and II error probabilities derived in Sections 7.3 and 7.4.

$$\begin{aligned} \tilde{P}_{\text{err II}}(f^*, k) &= 1 - q_{\text{II}} + q_{\text{II}} P(f_{\text{obs}} \sim \text{Binom}(n^{\text{lo}}, r^*s) < f^*) \\ \tilde{P}_{\text{err I}}(f^*, k) &= 1 - q_{\text{I}} + q_{\text{I}} P(f_{\text{obs}} \sim \text{Binom}(n^{\text{hi}}, rs) \ge f^* | r < r^*) \end{aligned}$$

and where $n^{\text{lo}}, n^{\text{hi}}$ are the lower and upper confidence bounds for n derrived from k described above. We also define the threshold function

$$\theta(k) = \max\{f^* | \tilde{P}_{\text{err II}}(f^*, k) \le \epsilon_{\text{II}}\}.$$

Since $\tilde{P}_{\text{err II}}$ monotonically increases and $\tilde{P}_{\text{err I}}$ decreases with f^* at fixed k, we can define k^* in terms of our maximal threshold function

$$k^*(r) = \min\{k | \forall k' \ge k, f^* = \theta(k'), \tilde{P}_{\text{err I}}(f^*, k') \le \epsilon_{\text{I}}\}.$$

We then observe that if the maximal threshold $\theta(k)$ satisfies $P_{\text{err I}} \leq \epsilon_{\text{I}}$, then the maximal threshold for any larger sample size $k' > k$ will at least satisfy the same type I error bound.

$$P_{\text{err I}}(\theta(k), k) \geq \epsilon_{\text{I}} \Rightarrow \forall k' \geq k, P_{\text{err I}}(\theta(k'), k')$$

It follows that

$$k^*(r) = \min\{k | \forall k' \geq k, f^* = \theta(k') = \theta(k), \tilde{P}_{\text{err I}}(f^*, k') \leq \epsilon_{\text{I}}\}.$$

Since $\theta(k)$ is not one-to-one, no inverse exist, but we can instead let

$$\theta^{-1}(f^*) = \max\{k | \theta(k) = f^*\}$$

and since $\tilde{P}_{\text{err I}}(\theta(k), k)$ strictly increases over the domain of k with fixed threshold $\theta(k) = f^*$,

$$k^*(r) = \min\{\theta^{-1}(\theta(k)) | \tilde{P}_{\text{err I}}(\theta(k), k) \leq \epsilon_{\text{I}}\}$$

Finally, if we only consider the domain where $k = \theta^{-1}(\theta(k))$, $\tilde{P}_{\text{err I}}(\theta(k), k)$ monotonically decreases in k, so finding the minimum k in this domain is equivalent to finding minimal $f^* = \theta(k)$

$$k^*(r) = \theta^{-1}(\min\{\theta(k) | \tilde{P}_{\text{err I}}(\theta(k), k) \leq \epsilon_{\text{I}}\})$$

Note that had we not narrowed the domain for k, a simple search for satisfactory values (e.g. with binary search) would be confounded by discrete jumps in $f^* = \theta(k)$, which cause $\tilde{P}_{\text{err I}}(\theta(k), k)$ to exhibit a sawtooth pattern that generally decreases while continuously increasing between discontinuous drops when f^* increments.

7.6 Evaluation by minimum sample size

Figure 4 depicts the minimum sample size of a perfect detector as a function of true contamination rate r as well as the sampling overhead incurred by detectors of imperfect sensitivities. These graphs assume plausible dynamics parameters with maximum contamination rate $r^* = 1/20,000$, $q_{\text{I}} = q_{\text{II}} = 0.99$, $\epsilon_{\text{I}} = \epsilon_{\text{II}} = 0.95$, and $\sigma_d = 50$. The relative sampling overhead cost of detectors with sensitivity 99%, 95% and 90% are below approximately 2%, 6% and 12% respectively, which we expect to be manageable.

8 DISCUSSION & CONCLUSIONS

We demonstrate that modern image modeling techniques can contribute to making SIT a viable approach for mosquito population control by efficiently validating that female releases are negligible. Our experiment results and sample size calculations suggest that achievable detection accuracies are sufficient for making statistically justified decisions with respect to a factory's contamination rate while incurring minimal sampling overhead. Furthermore, the fact that these calculations neglect the advantage of predicted counts for each plate hints that predicted counts add little value to the contamination rate measurement.

Experimenting with context region configurations leads us to believe that the context region is worthwhile, but the corresponding mask adds little-to-no value. Discarding this feature enables the possibility of implementing our architecture as a fully convolutional network [15] thus avoiding redundant convolutions, which are particularly prevalent given context region overlap.

Additional experiments with focus and interpolation sizes reveal that detection sensitivity is better with the smaller focus size while counting accuracy prefers larger patches. Similarly, while counting accuracy performs best with small focus size detections sensitivity is optimized with single point representation. These experimental results roughly correlate to our analysis of theoretical limits of sensitivity under patch-based insect scoring. In most cases, measured specificity was within one or two percentage points from that of optimal binary scoring.

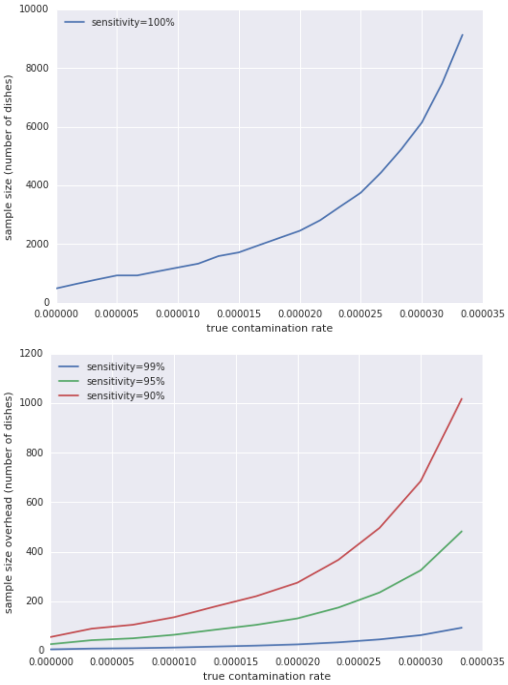

Figure 4: Above: Upper bound on the minimum sample size required to satisfy type I and II error requirements with a perfect detector model. Below: Sampling overhead incurred by using imperfect counters are varying sensitivities to meet the same set of type I and II error requirements.

We also examined the effect of patch-based detection on optimal per-mosquito sensitivity and specificity and observed that a finer-grained prediction target could lead to tighter localizations with fewer false positives.

Finally, we note that careful analysis of our application's requirements led to unexpected observations with implications for modeling approaches and metric interpretations.

ACKNOWLEDGMENTS

The authors would like to thank Dale Webster and Nathan Silberman for providing early modeling advice and direction, Yang Feng and Brian Patton for reviewing and advising the development of early statistical models, Ashwin Kakarla and Aparna Shorapurkar for assisting with the data-labeling process, and Michael Terry for reviewing and editing this paper.

A EXPERIMENT DETAILS

The context region and focus size / interpolation size studies trained under differing conditions detailed below.

A.1 Context Region

The context region study ran training for 4 million steps and lacked learning rate decay, batch normalization, and weight decay.

A.2 Focus size and interpolation size

This study applied all of learning rate decay (with decay schedule of 5% per 10K steps), batch normalization and weight decay but only trained for approximately 1 million steps per model with some variation between models. We list approximate final iteration numbers in the table below.

	$R = 200$	$R = 300$
$M = 1$	1.2M	1.1M
$M = 2$	1.3M	1.1M
$M = 10$	1.3M	1.1M

The experiment also applied a rectified linear activation to model predictions as counts should be non-negative and we expected this final activation to make modeling a distribution with mostly zero values easier to do.

Some configurations yielded somewhat miscalibrated counting predictions, which may suggest a need to tune learning rate hyperparameters.

Subsequent experiments suggest softmax classification yields more accurate count predictions.

REFERENCES

[1] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Luke Alphey, Mark Benedict, Romeo Bellini, Gary G Clark, David A Dame, Mike W Service, and Stephen L Dobson. 2010. Sterile-insect methods for control of mosquito-borne diseases: an analysis. *Vector-Borne and Zoonotic Diseases* 10, 3 (2010), 295–311.

[3] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. 2016. Counting in the Wild. In *European Conference on Computer Vision*. Springer, 483–498.

[4] Danilo O Carvalho, Derric Nimmo, Neil Naish, Andrew R McKemey, Pam Gray, André BB Wilke, Mauro T Marrelli, Jair F Virginio, Luke Alphey, and Margareth L Capurro. 2014. Mass production of genetically modified Aedes aegypti for field releases in Brazil. *JoVE (Journal of Visualized Experiments)* 83 (2014), e3579–e3579.

[5] Prithvijit Chattopadhyay, Ramakrishna Vedantam, R. S. Ramprasaath, Dhruv Batra, and Devi Parikh. 2016. Counting Everyday Objects in Everyday Scenes. *CoRR* abs/1604.03505 (2016). http://arxiv.org/abs/1604.03505

[6] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1. IEEE, 886–893.

[7] David A Dame, Christopher F Curtis, Mark Q Benedict, Alan S Robinson, and Bart GJ Knols. 2009. Historical applications of induced sterilisation in field populations of mosquitoes. *Malaria journal* 8, 2 (2009), 1.

[8] VA Dyck, J Hendrichs, and AS Robinson. The Sterile Insect Technique: Principles and Practice in Area-Wide Integrated Pest Management. 2005. *Dordrect, Netherlands: Springer* (????).

[9] Geoffrey French, Mark Fisher, Michal Mackiewicz, and Coby Needle. 2015. Convolutional neural networks for counting fish in fisheries surveillance video. (2015).

[10] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1725–1732.

[11] Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[12] Edward F Knipling. 1959. Sterile-male method of population control. *Science* 130, 3380 (1959), 902–904.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[14] Victor Lempitsky and Andrew Zisserman. 2010. Learning to count objects in images. In *Advances in Neural Information Processing Systems*. 1324–1332.

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

[16] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.

[17] Thomas Moranduzzo and Farid Melgani. 2014. Automatic car counting method for unmanned aerial vehicle images. *IEEE Transactions on Geoscience and Remote Sensing* 52, 3 (2014), 1635–1647.

[18] T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. 2016. A Large Contextual Dataset for Classification, Detection and Counting of Cars with Deep Learning. In *European Conference on Computer Vision*. Springer, 785–800.

[19] Philippos A Papathanos, Hervé C Bossin, Mark Q Benedict, Flaminia Catteruccia, Colin A Malcolm, Luke Alphey, and Andrea Crisanti. 2009. Sex separation strategies: past experience and new approaches. *Malaria journal* 8, 2 (2009), S5.

[20] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229* (2013).

[21] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *CoRR* abs/1512.00567 (2015). http://arxiv.org/abs/1512.00567

[22] World Health Organization 2015. 10 facts on malaria. http://www.who.int/features/factfiles/malaria/en/. (2015). [Online; accessed 31-October-2016].

[23] World Health Organization 2016. Dengue and severe dengue. http://www.who.int/mediacentre/factsheets/fs117/en/. (2016). [Online; accessed 31-October-2016].

[24] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. 2015. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 833–841.