UNIVERSITY OF CALIFORNIA
SANTA CRUZ

CLASSIFICATION OF THE SEX OF DROSOPHILA SUZUKII WITH PRE-TRAINED NETWORKS

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PHYSICS

by

Drew Marcus Bischel

June 2023

The Dissertation of Drew Marcus Bischel
is approved:

Professor Joshua Deutsch, Chair

Professor William Sullivan

Professor Anthony Aguirre

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Classification of the sex of Drosophila suzukii with pre-trained networks

by

Drew Marcus Bischel

There has been a recent trend to applying deep learning methods compared to shallow methods for automatic identification of insects. Classification strategies built around algorithms with deep learning architectures at their center like YOLO and others require large amounts of data to making learning successful and are often augmented with tens of thousands of images or more to achieve excellent performance. Recent pre-trained models of deep neural networks have significantly reduced the amount of data required to create accurate classification algorithms by ingesting and training on a huge data set different than the target task and using the resulting encoding to transfer information to a new task. This work shows that recent performance gains from models pre-trained on huge data sets are effective as image encoders for the classification of the sex of spotted wing drosophila (SWD). A data set of 676 SWD microscope images is created to evaluate classification models for use in automation of the sterile insect technique (SIT), which requires large amounts of male SWD to be identified and separated. Binary classification models trained on top of image encoding from new models based off of visual transformers [3] pre-trained on over 400 million images with CLIP [2] are able to achieve accuracy as high as 96.7% when trained with LogReg and similar classifiers on augmented data from the SWD image set.

Other models pre-trained on the ImageNet data set of 14 million images also performed well, approaching 92% with VGG models and 90% with MobileNetV2 model. Image segmentation of the data set is then investigated as a source of corroboration for the identification of the morphological features responsible for classification, and an out-of-distribution data set is collected to evaluate classification and segmentation results on more diverse and difficult examples. While robust identification of features special to SWD remains, classification accuracy is not a guarantee on data which differs substantially from the factory or laboratory setting on which it is trained and additional data may be needed for training on use-cases outside of SIT such as for applications on the farm or for automated identification in insect traps. This emphasizes a fact which is not elaborated on for many insect detection models in the literature: that their models are not likely robust in situations where the data is significantly OOD and for situations which may not be adequately covered without specialized augmentation methods or additional data. Nonetheless results indicate that pre-trained models have advanced to the point where they can play a central role in securing the food supply from potentially billions of dollars of damages every year from pests such as SWD.

Figure 0.1: A colorful image taken during a family reunion.

I dedicate this

to my mom and dad,

who take up a very big space in my heart.

Acknowledgments

First I would like to thank my PhD advisor Josh, for sharing thoughts, ideas and offering insightful feedback whenever I needed it, and even more for sharing this experience in an official and unofficial capacity during my association with the university. Second I would like to thank my defense committee for agreeing to read this, Bill for suggesting the idea of using to machine learning to classify Drosophila suzukii and to Anthony for acting as a co-advisor and educator to me throughout my PhD.

I am grateful for my experience in academia, which has culminated so far in my writing of this thesis. There are far too many teachers and coaches to acknowledge throughout my learning and work in educational institutions from K-12, to my undergraduate at San Francisco State University, and now my PhD at UC Santa Cruz. I am grateful to the physics department at UCSC for supporting me through a period of time longer than I have spent at any institution. All of the professors I've been TA for have been invaluable as mentors, especially those who I was privileged to work with closely in Phys 133 and Phys 134, and to George Brown who was essential in curating so many aspects of my educational experience for these lab classes. I'm thankful for my friends and office mates, like Dana, Andrew, Chris, Dominic, David, Johnny and Arturo were always delighted to tell me about their research. Of course I could go on about all my other class mates who I had the opportunity to get to know. All of my colleagues here have been wonderful friends that I've gotten to learn so much from. Friends from years past, including Aren, Steve, Greg and more were important mentors and collegues to me. All of the graduate advisors as part of the department staff including Cathy, David, Ben, Amy

and Jared were essential in coordinating deadlines, scheduling, TAships and moral support.

I would like to thank my mom and dad, being the ones who this thesis is dedicated to, for raising and loving me and supporting me through many happy and sad moments in life. My brother Tyler and sisters Mandy and Heather have all been inspirational throughout my life in numerous capacities.

Chapter 1

Introduction

1.1 Motivation

Drosophila suzukii Matsumura, commonly referred to as spotted wing drosophila (SWD), is an invasive species of fruit fly responsible for inflicting significant damage to the global supply of soft and stone fruits such as strawberries, blackberries, raspberries, cherries, blueberries, grapes, nectarines, peaches, apricots and others. The economic impact can be severe; some estimates indicate that there is potential for over $700M worth of crop damage due to SWD every year in the United States alone [7, 8, 9]. D. suzukki is capable of causing such widespread crop damage for a number of reasons. The female has a serrated ovipositor whose insertion causes physical damage to the fruit and very often leads to secondary infections from other insects and pathogens within the oviposition wound. Eggs deposited in the fruit eventually turn into larvae which causes the fruit to soften and rot. Compounding this, SWD has a tendency to target fruits in early growth stages before they are ripened, resulting in significant damage before they can be harvested. D. suzukii represents a challenge to existing integrated pest management (IPM), as the species

has shown resistance to some insecticides and has multiple generations within a season, further reducing the viability of traditional insecticides, which themselves can have a negative impact on biological control agents and leave chemical residues on fresh produce [10, 11].

Sterile insect technique (SIT) is an environmentally conscious method which has shown success in the past at suppressing outbreaks of many insects. SIT functions by releasing sterile insects that reduce the population by competing for mates. Sterile insects are not capable of self-replicating. Integrated with additional control strategies, the SIT has succeeded at thwarting insect pests in high-profile situations like the Mediterranean fruit fly, Mexican fruit fly, melon fly and others [12].A key component of developing a system that is able to facilitate SIT is one which identifies and separates males from females so that sterilization may be induced one gender; in the case of SWD the males are the target to be collected. Studies on the capability of machine learning to identify between male and female specimen of Drosophila suzukii would contribute to the food grower community's effort to manage this pest. The main purpose of this work is to procure a data set capable of being modeled to classify male SWD specimen with fidelity and to demonstrate that pre-trained networks are suitable candidates to develop industrial classification applications for SIT and other entomological applications.

In addition to industrial scale rearing operations, monitoring and surveillance equipment could be utilized in traps on the farm itself. The monitoring of traps is considered in integral part of SWD IPM [13]. Public or private land spaces that keep tabs of local insect populations would benefit from availability of such technology.

The World Health Organization (WHO) identified entomological equip-

ment as an important pillar of focus for research needed to address the global threat of vector-borne diseases that affect humans [14]. Drosophila in particular are often used as specimen for research in genetics, which in many circumstances utilizes highly skilled professionals who can identify morphological differences among generations. Dissemination and application of image classification technology may be helpful in identifying species with targeted traits or by detecting morphological mutations over generations of fruit flies. While the primary focus of this work is placed on automated study of SWD morphology for applications in crop production and food health, there could be relevant overlap with other entomology and health projects currently in progress. The bounty of data that entomological data sets offer may help feedback into the machine learning community to contribute to advances in detection technology related to other biological data sets which may have an impact in medicine. Therefore it is worthwhile to produce data sets that improve the community's ability to identify morphological characteristics. With recent advances in capability of machine learning to consume and analyze large amounts of data, regular evaluation of new technologies is needed in industry in order to improve sectors such as health, food security, medicine as well as the broader economy.

1.2 Preliminaries

A small amount of background material is helpful to introduce some of the concepts related to the motivation and the methodology in later sections. First a brief summary of the spread of Drosophila suzukii to Santa Cruz county and the pacific northwest, along with some information about the usage of sterile insect technique to manage this species are discussed. Then a brief introduction of some

of the terminology and concepts used in the methodology for image classification is followed by a relevant though non-exhaustive review of entomological machine learning applications and their classification results before transitioning into the methodology used in this work.

1.2.1 Historical notes on the spread of SWD in the Pacific Northwest

The spotted wing drosophila (SWD) is a pest that causes damage to soft-skin fruits such as cherries which has in recent years established itself in the continental United States, Canada and Europe. Though the earliest records of this species originate from Japan in 1916 and are described by Matsumura from Japan in 1931, the species has also been observed in Korea, Thailand and India. Evidence of dispersion of SWD was gathered in Oahu, Hawaii in 1980 and afterwards in other Hawaiian islands [15]. Observations of SWD in the US mainland were first reported in August 2008 in Santa Cruz County in California, and the following year in May additional infestations were reported along the central coast of California in cherry orchards [7, 15]. The rapid spread of this serious pest was of heightened concern to local farmers; by 2009 the presence of this species had expanded to over 20 counties in California from San Diego all the way up to Humboldt County [15]. Subsequent trapping efforts confirmed the presence of SWD in Washington and Oregon, as well as other observations in British Columbia and even Florida [15, 7, 16]. By the time of its detection, eradication of this invasive species was considered impossible by the California Department of Food and Agriculture [7]. Through 2010 to 2012 the fly continued to spread into the eastern states and is likely observable in many more states.

1.2.1.1 Economic concern of SWD

The disadvantageous effects of SWD on local economies are primarily concerned with raspberry, blackberry, cherry, blueberry crops, and especially with the potential to damage central California's large strawberry production; however, this species is adaptable to a wide range of non-agricultural hosts, which enables its persistence in woodland habitats that often contain wild blackberries and cherries in the summer as well as other fall-bearing fruits. [17, 18, 8]. The yield loss estimates for 2008 when damage was first noticed vary from negligible to 80%, depending on which crops and locations were most affected. One estimate from 2015 pins the annual economic loss to growers at $ 700 million [9]. Economic losses are expected to fall as producers learn to manage SWD more effectively though new control tactics [19].

1.2.1.2 Notes on the biology of SWD

D. *suzukii* has a brown to yellow abdomen with dark bands and strong red eyes. Males have a single dark spot on the tip of each wing, which is an potential target for any image based classification algorithm that is designed to distinguish gender, since the females do not have the spot; however, the orientation of the fly in the image may restrict the amount of pixels and lighting conditions necessary to see the spot. The foreleg of the male has sex combs that look like dark bands on the first and second tarsi,and they are much more difficult to see at low magnifications than the male's spotted wing at the wing tips. The female has a visible serrated ovipositor which is the most important distinguishing factor in comparison to the males.

1.2.2 Management of SWD with SIT

1.2.2.1 Sterile Insect Technique

The sterile insect technique (SIT) is an approach to biological insect control, where sterile males compete with fertile males to mate with the females. Since the sterile males are not capable of self-replicating, the females that they mate with cannot deliver offspring. The effects of X-ray photon irradiation are applied to the insects' reproductive cells in order to induce sterilization. A large amount of sterile males are released into the targeted area, where the population decreases accordingly. The advantage of SIT compared to other IPM techniques is that it does not leech chemical pesticides into the regional environment and it does not introduce non-native species into an ecosystem; however, SIT may take repeated applications. SIT has been successfully implemented to manage outbreaks of the screw-worm fly, which caused annual losses to the meat and dairy supplies in the 1950s, and successes have also been demonstrated on fruit fly pests such as the Mediterranean fruit fly, the melon fly, and the Mexican fruit fly, as well as other insect species.

Estimates taken from multiple operational SIT programs and rearing facilities indicate that from 1963 to the present over a trillion sterile insects have been delivered in trans-boundary shipments. While they mostly consist of the Mediterranean fruit fly, they include fruit flies, screw-worm flies, tsetse flies, moths and mosquitoes. Protection of horticultural and livestock industries through reductions in crop and livestock losses, as well as improvements in access to high value markets without quarantine restrictions have been benefits of using the technology. The return on investment is further raised by protecting the environment from disruptive invasive species and reducing human health and environment costs of frequent

insecticide use [12].

1.2.2.2 Studies of SIT for Drosophila *suzukii*

SIT has been successfully implemented via a proprietary method, suppressing wild female D. *suzukii* by up to 91% compared to the control suites with sustained and dynamically targeted releases of sterile males in strawberry-growing open polytunnels [20]. A study of irradiation doses to sterilize D. suzukii found that females irradiated with 50 Gy or more had almost no fecundity, while experiments irradiating males with 120 Gy decreased egg hatch rate exponentially [21]. Another study attempted to lower the amount of radiation required by combining Wolbachia symbiosis, which has sterilization effects in its host flies, with the irradiation method [22]. Still other pest management strategies which aim to disrupt the reproductive cycle are actively researched. For example, gene-driven approaches for the suppression of pests which favor biased inheritance from generation to generation have been studied [23]. See [24] for a detailed overview of SWD management strategies and information.

1.2.2.3 Role of image classification in SIT

One of the major challenges identified for the deployment of SIT is the automation of as many parts as possible of the process of rearing, sex separation, irradiation, handling, packaging and releasing sterile specimen [25]. In order to separate male from female specimen a classification algorithm can be used to automatically identify which sex the image belongs to and then send a command for processing of the insect. It may be necessary to extend this application to other aspects of the handling or irradiation processes, depending on the details of the SIT

operation. For example, continuous surveillance of separated populations may be desirable to ensure that population of one sex (e.g. females) does not contaminate the other (males). Classification algorithms developed in the cloud could be accessed by field techs to identify presence of insect populations to be targeted with SIT. Auxiliary surveillance objectives may be needed to identify the absence or presence of a target on a slide, or the location of specimen in a chamber.

1.2.3 Supervised machine learning

The field of machine learning is a well-developed area of focus with many decades of fundamental and ground-breaking research stemming from ideas in computer science, mathematics, physics and engineering. In the last decades, in large part due to increases in availability of memory and compute, we have seen a tremendous growth of research demonstrating progress in the performance of machine learning algorithms. One paradigm of machine learning called supervised learning focuses on problems where each data point made available to the algorithm is in the form of an input feature and an associated output label which forms the supervisory signal. The goal of the algorithm is to predict the output label when given unseen input features. The performance of the algorithm is measured through some sort of generalization error procedure used to validate its accuracy.

There are numerous challenges that are typically considered for supervised learning algorithms. Among these are the complexity, size, balance and distribution of the training data, dimension of input, pathological redundancy or noise in the data set, flexibility and robustness of the algorithm to out-of-distribution data, tendency of the algorithm to overfitting or excessive bias, and many others. These issues can be addressed within the model or architecture itself or by exhaustive amounts of

new or augmented data.

1.2.4 Recent big data approaches on images and text

1.2.4.1 ImageNet

The ImageNet database is a collection of more than 14 million images which have been hand labeled in more than 20,000 categories. Fortunately for this work, this data set includes a variety of animals, plants, insects, transparent man-made objects and other items seen in photographs across the internet, which should be relevant for insect classification. This resource was created to establish a clear community goal for improvement in AI as well as to meet demand for the large amount of data needed to enable generalization among machine learning methods [26], and forms the basis of the ImageNet Large Scale Visual Recognition Challenge [27]. It is often used as a benchmark and pre-training module for initializing weights in large neural networks and other image consuming models. In the case of this work it is used to initialize the weights of the VGG and MobileNet models.

1.2.4.2 Semi-supervised and unsupervised learning

In recent years pre-training methods have advanced the way we consume, analyze and utilize large amounts of data. One of the key advantages of pre-trained networks is that they are reusable for a wide variety of tasks; one no longer needs to train a network to extract the elementary features of an input, which reduces the amount of data required to achieve performance gains.

Particularly, advances in natural language processing (NLP) have shown that task-agnostic objectives such as masked language modeling can be scaled across many orders of magnitude in model capacity, data consumed and training compute

to achieve continued performance gains [28, 29]. Generative language models like GPT-3 trained at such large scales have exhibited zero or few-shot generalization, demonstrating an impressive range of capabilities whose designs were not directly intentional [30]. Further progress on this front is using image input to augment and improve linguistic understanding from AI models which generate text output using image and text inputs [31]. One recent set of models that use natural language text paired with images combines them into a separate image and text encoding that will be used in this work. They capture a diverse set of training material at a large scale, collectively called CLIP, learned from a data set of 400 million (image, text) pairs [2].

1.2.5 Related work

A review of studies of image capture and classification shows a recent trend toward applying deep learning methods compared to shallow learning methods for automatic identification of insects [32].

A data set of fruitfly species *Bactrocera Zonata* and *Bactrocera orsalis* consisting 2000 images was used to train YOLOv5 for species classification and reported 85% accuracy [33]. Another work produced a training set of over 10,000 images from microscope slides, augmented to over 190,000 images, of various mosquito species was used to train models with high precision and sensitivity (reported 99% and 92.4% respectively) in order to simultaneously localize and classify species of the gender of field-caught mosquitoes with a focus on the multi-stage use of the YOLOv3 algorithm [34]. More work investigates training algorithms combining custom CNNs with SVMs, random forests and other models to carry out classification of 760 fruit fly images of four different categories of Bactrocera species images with complex

backgrounds, with accuracy results of 92% for CNN+SVM architectures [35].

One recent publication demonstrated promise in determining the physiological age of the eggs of two *Tephritid* fruit fly species with accuracy of between 75% and 83.16% depending on the species with 892 images (augmented to over 35,000 images) using the Inception v1 algorithm [36].

High resolution wing and aculeus images from species of the genus *Anastrepha* have been used in the development of automatic identification of fruit flies were able to achieve high accuracy by fusing multiple learning techniques (combining decision tree, k-nn, SVM, and other approaches), whose objectives included studying the morphological relevancy of certain regions of the fruit fly during classification of species [37]. A follow-up to their work studied deep feature-based classifiers for fruit fly identification and found that VGG16 and VGG19 achieved high accuracy (95.68% and 94.34%) on their data set of 301 2560x1920 images of three categories of fruit fly species [38]. The work of this thesis has overlap with theirs, as the male wing and female serrated ovipositor are identified as important regions of classification for SWD; however, the resolutions used for classification in this work are much smaller.

1.3 Methodology

In this section the algorithms underlying the image classification used in later chapters are introduced. First an overview of the pre-trained networks used as an essential component of image classification is given, then the concepts behind the probe classifiers used for binary classification are discussed, followed by the framework used for image segmentation. Lastly the performance metrics, validation

and uncertainty concepts are introduced which will help convert the classification results into an interpretable form which are a subject of analysis in the later chapters.

1.3.1 Pre-trained networks

Collecting data for a classification task can be tedious and expensive. The data can occur with low frequency or be difficult to find, and the proper imaging or data observation equipment must be carefully selected beforehand so that the data set will be applicable to the environment where the algorithm will be employed. Pre-trained networks have proved invaluable for transfer learning, where features learned from a large data set independent from the task data set can be used in conjunction with the task data set. There are numerous methods for training these image encoders from classification data, which are used as a base which can either be fine-tuned to the target data set or used as input to an additional network trained on the specific classification data. Advancements in transfer learning have allowed the amount of data required for deep learning to be successful to be reduced significantly. While this is exactly what should be expected in the future of advancement in AI (the effort to succeed at your task will be made progressively smaller), it is nonetheless remarkable that we are able to routinely produce algorithms comparable to or better at classification than a human with relatively small amounts of training data.

1.3.1.1 ImageNet-based pre-trained models

The Visual Geometry Group at Oxford produced pre-trained models which became highly recognized and subsequently used in many classification applications. The VGG net models, VGG16 and VGG19, with 16 and 19 layers of small 3x3 ConvNet models come from submissions to the ImageNet Challenge 2014 competition

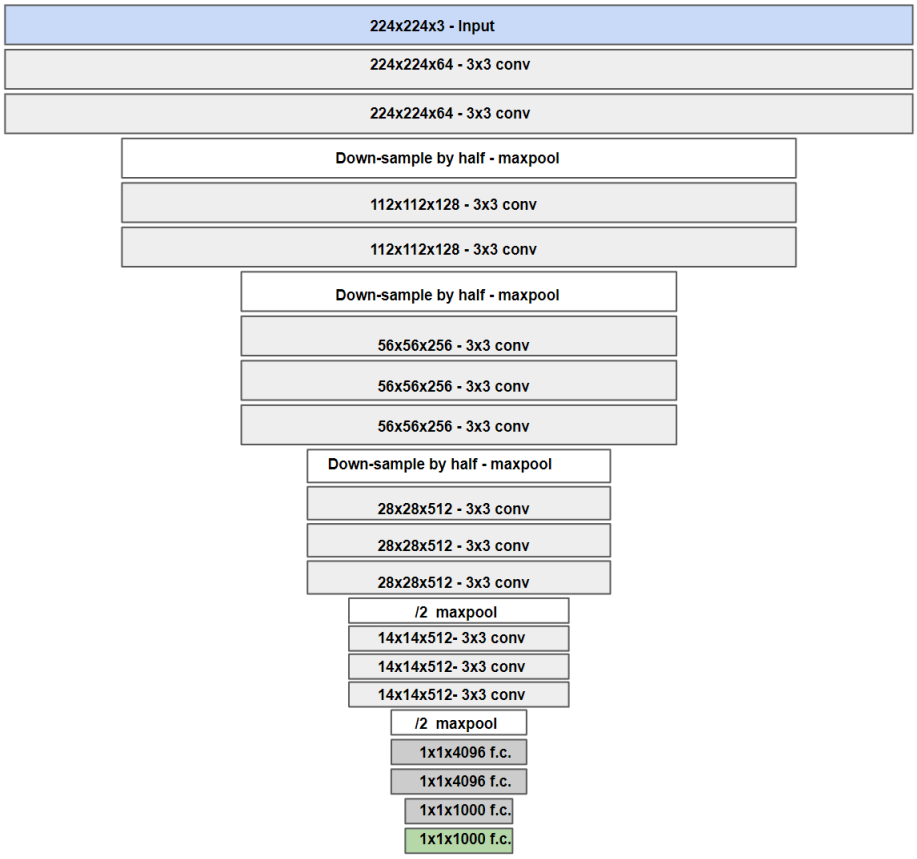

Figure 1.1: This is an example depiction of a VGG16 network. VGG19 has an additional layer on each of the last three blocks of conv layers. Input as shown goes sequentially from top with output on the bottom with fully connected layers.

where they achieved first and second prize in localization and classification respectively [39]. VGG16 and VGG19 had approximately 138 million and 143 million parameters respectively, though it was found that about 100 million could removed from the fully connected layer without losing performance. The goal of the VGG models was to increase classification accuracy by increasing the depth of ConvNets.

VGG net has an input size of 224x224 RGB images, and passes its input through either 16 or 19 3x3 convolutional layers of stride-1, with five layers of max pooling each downsampling between the convolutional layers. Behind this there are three fully-connected layers followed by a softmax layer.

The MobileNets series of models are often used for tasks where a relatively

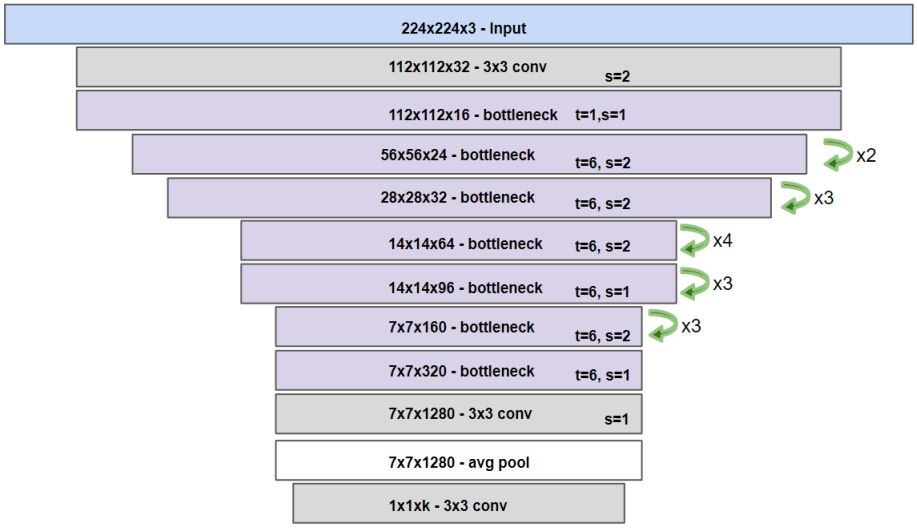

Figure 1.2: This is an example depiction of a MobileNetV2 network. MobileNetV2 uses depthwise convolutions to improve efficiency in collecting image features. Input as shown goes sequentially from top with output on the bottom with fully connected layers. Each convolutional layer has either a stride s=1 or stride s=2 depthwise convolution layer, and repeats a number of times (green arrows). Each residual bottleneck layer has an expansion factor t which increases the number of channels output by the Relu 1x1 conv layer at the start of the bottleneck (as described in [1]

)

efficient model is needed at minimal performance loss. Specifically, this work uses an implementation of MobileNetV2 [40] trained on ImageNet data. It distinguished itself from previous models by using lightweight depthwise convolutions in its intermediate layers, which apply a single convolutional filter per input channel.

1.3.1.2 CLIP models

The CLIP models used in this work are composed of an image encoder and a text encoder that are used to incorporate visual information with text information under joint training (see for example p.5-37 in [2]). The text encoder is a transformer architecture [41] with modifications reported in [42]. The base size

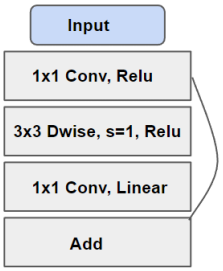

Figure 1.3: This is a stride-1 type bottleneck layer used in MobileNetV2. An input with k channels is expanded with expansion factor t to $(t \times k)$ channels with the first operation, and after the depthwise convolution the output becomes $\frac{h}{s} \times \frac{w}{s} \times (t \times k)$ for an input with height h and width w. There is a residual connection after the 1x1 Conv-Relu operation added to the output of the final 1x1 linear conv operation.

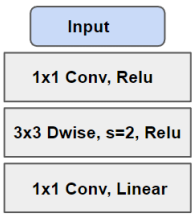

Figure 1.4: This is a stride-2 type bottleneck layer used in MobileNetV2, which works similarly to the stride-1 block except that the stride reduces the output image size by half and there is no residual connection after the 1x1 Conv-Relu operation.

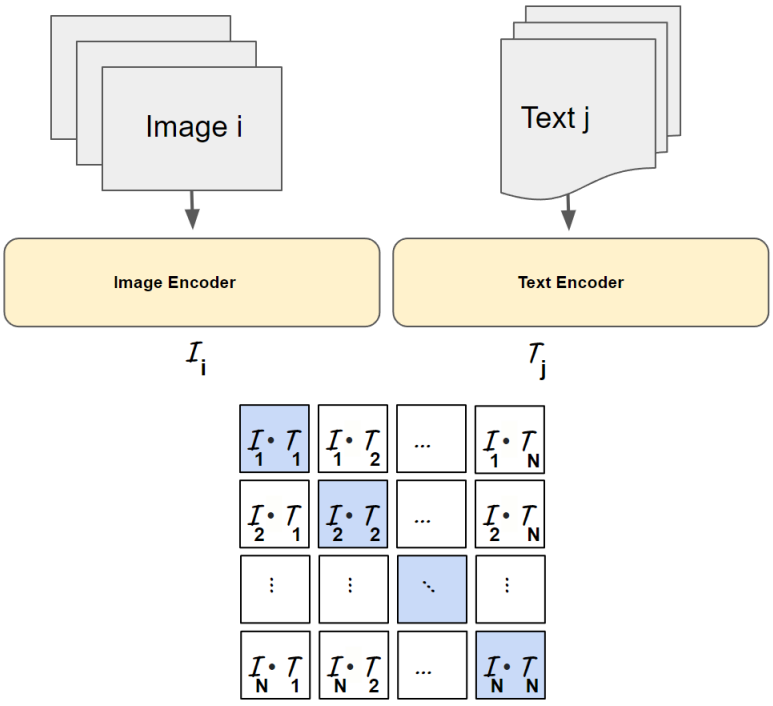

Figure 1.5: CLIP uses contrastive learning, which utilizes both positive examples and negative samples such that the loss function maximizes the distance between negative examples and minimizes the distance between positive examples. As shown there are $N^2 - N$ negative examples and N positive examples. CLIP models are trained on a dataset of 400 million (image, text) pairs and uses contrastive objectives to learn image representations from text. See [2] for a full description of their work.

for their Transformer uses a 512-wide model with 8 attention heads stacked into 12 layers with 63M parameters. While the text encoder is not explicitly used in this work, it could be useful in future work, especially if a multiple classifiers are used in conjunction on top of a single pre-trained CLIP model for instance to tell if the slide is empty or has a fly in it.

The image encoders for CLIP are developed in two main variants, one is a ResNet variant and the other models are based off of visual transformers, specifically ViT-B/32 and ViT-B/16 from [3]. Two of the ResNet variants, RN50x16 and RN50x4 are modified from [6] in the EfficientNet style, while RN101 and RN50 are

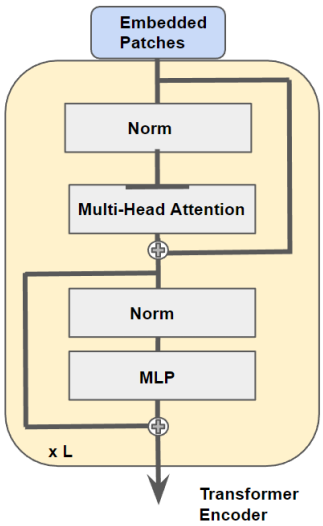

Figure 1.6: This is a depiction of the VIT transformer encoder layer described in [3].

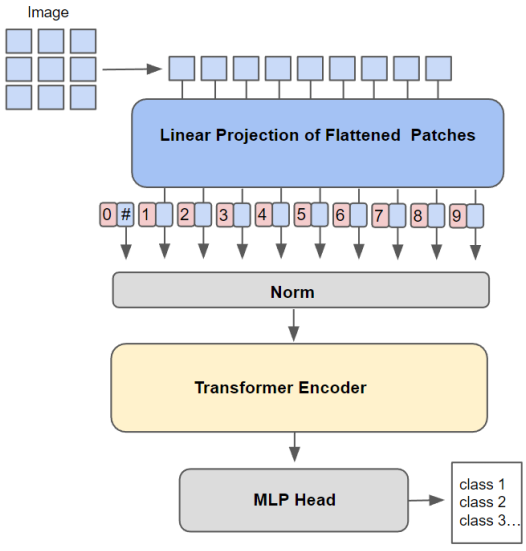

Figure 1.7: The implementation used for CLIP follows the original implementation but with an additional batch normalization layer applied on the output of the combined image patch-position embeddings.

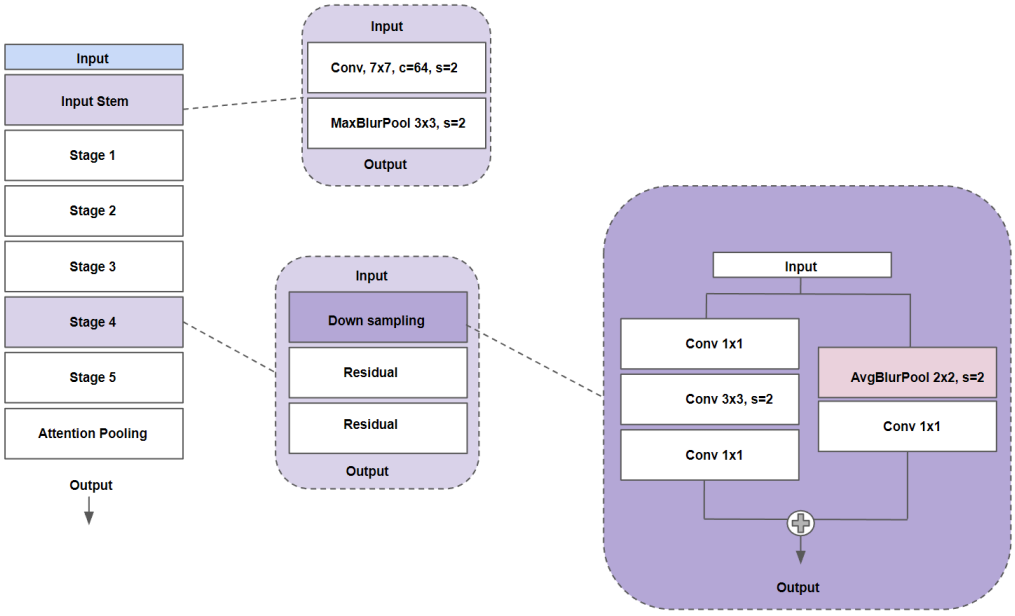

Figure 1.8: This is a depiction of the modified ResNet-D with some example portions modified to reflect the improvements from [4] in the down-sampling and improvements in anti-aliasing and equivariance from [5]. RN50 and RN101 use versions of this type of model, while RN4 and RN16 use versions of efficientNet from [6] which scale depth and width of convolutional layers at a fixed ratio as the network deepens.

described in [2] modified from [4]. According to [2] the ViT transformers are about 3x more compute efficient than the CLIP ResNets.

1.3.2 Binary classification

Binary classification refers to the task of identifying members of a population from two classes, which in the case of this work is the task of separating SWD into male and female classes. There are many algorithms developed to carry out binary classification and their applicability depends heavily on the complexity or difficulty of the problem to be solved. For image classification it is often necessary (as it is in this work) to encode the images into a condensed feature representation which is useful for binary classification. This is done with the pre-trained models, however an additional decision must be made which specifies how the image encod-

ing will be converted into a binary positive or negative identification. Other neural network architectures may be required to use the image embedding effectively for generative tasks which are more complicated than binary classification, but a linear model is sufficient to carry out binary classification from the pre-trained embeddings for the purposes of this work, as it is most convenient to implement. A linear model is one which makes a classification decision based on a linear combination of the elements of the feature vector. They are often preferred over non-linear classifiers because they are easier to use and take less time to train.

A logistic regression classifier (also known as a MaxEnt or Logit classifier) is a linear classifier which is a common choice for carrying out binary classification, where the prediction probability is estimated with a logistic function $p(\vec{x}) = \frac{1}{1+e^{-\vec{\beta}\cdot\vec{x}}}$. Where linear regression seeks to minimize the squared error loss between the outcome of the k-th data point y_k and the prediction for that output p_k, logistic regression minimizes log loss for the k-th data point with the function $-\ln p_k$ if $y_k = 1$ and $-\ln(1-p_k)$ if $y_k = 0$. This can be combined into a single expression called the cross entropy of the predicted distribution from the actual distribution where the loss $L = \sum_{k=1}^{K} -y_k \ln p_k - (1-y_k)\ln(1-p_k)$. This sum is the negative log-likelihood which when minimized is equivalent to maximizing the likelihood function itself, so this process is also sometimes referred to as maximum likelihood estimation which maximizes the probability that a particular logistic function models the data set.

Another linear classifier used in this work is the linear support vector machine. The idea behind a support vector machine is to construct a hyperplane in feature space that separates positive examples from negative examples with a maximum margin. In this process the input feature vector $\vec{x}_i$ is assigned either +1 or -1 depending which class it is in. A hyperplane can be defined as the set of points

$\vec{x}$ which satisfies $\vec{w}^\top \vec{x} - b = 0$, where $\vec{w}$ is a vector normal to the hyperplane and $b/\|\vec{w}\|$ is the offset of the hyperplane from the origin along $\vec{w}$. For training data that is linearly separable a hard-margin can be defined so that a feature vector $\vec{x}_i$ where $\vec{w}^\top \vec{x}_i - b \geq 1$ is assigned to the +1 class and a feature vector satisfying $\vec{w}^\top \vec{x}_i - b \leq -1$ belongs to the -1 class. This can be written as an optimization problem that minimizes the L2 norm of the square of the normal vector to the hyper plane $||\vec{w}||^2$ over $\vec{w}$ and b subject to the constraint that $y_i(\vec{w}^\top \vec{x}_i - b) \geq 1$, where $y_i = \pm 1$ is the class of the data point. This is extended with a soft-margin for cases where the data are not linearly separable by using the hinge loss function $L_{hinge} = \max(0, 1 - y_i(\vec{w}^\top \vec{x}_i - b))$ and a regularization parameter $\lambda > 0$ to minimize $L = \lambda \|\vec{w}\|^2 + \frac{1}{K} \sum_{k=1}^{K} L_{hinge}$ over all K training examples. The regularization parameter λ determines the flexibility of the margin size versus requiring that all x_i belong on the correct side of the margin. This can be further extended by replacing every dot product with a nonlinear kernel function, which may result in non-linearity of the original input space but still uses a hyperplane in the transformed feature space. A kernel to use for example is an inhomogeneous polynomial kernel $k(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j + r)^d$. An advantage of SVMs is that they are effective in high dimensional spaces where the number of feature dimensions is greater than the number of samples; however, if the number of features is much greater than the number of samples careful selection of kernel functions and regularization are needed to alleviate over-fitting.

1.3.3 Image segmentation

In the previous section binary classification was discussed and a few common algorithms were outlined that are able to distinguish between positive and negative classes. These algorithms have various extensions into the realm of multi-

class classification, where instead of positive and negative examples there are more than two classes that the algorithm is trained to distinguish between. In particular the sparse categorical cross-entropy loss is used to carry out multi-class classification. Formally the cross entropy between the true and predicted distribution can be defined in terms of the Shannon entropy and the Kullback-Leibler divergence to quantify the average extra number of bits required in excess of the Shannon entropy for events from the true distribution to be encoded by the predicted distribution.

Multi-class classification has many applications; one such application is image segmentation, where each pixel is assigned a class prediction which results in the output of an image mask containing the locations on the image classified into one of multiple exclusive categories. While informally the image segmentation in this work is still used to carry out binary classification a number of additional class possibilities are included which distinguish the fruit fly from the background as well as pixels belonging to the edge of the male or female fruit fly bodies.

The added complexity of pixel-wise predictions and multi-class classification require additional architectures that are useful for transmitting information both across an image and across smaller subsections of an image. U-Net [43] is a convolutional architecture which was originally developed for biomedical image segmentation that achieved great success in this domain. It has been adapted to many tasks and is a widely used template for machine learning projects today which carry out tasks such as brain and liver image segmentation, protein binding site prediction, medical image reconstruction and other image-to-image translation tasks.

U-Net is built from a down-sampling and up-sampling structure where repeated applications of convolutions followed by rectified linear units and max pooling operations are used as a bottleneck similar to the architectures discussed in previous

sections which successively reduce spatial information into an image encoding. During the up-sampling portion of the network where the resolution is increased from the base of the image encoding in order to eventually provide a full image mask, higher resolution network levels are concatenated from the down-sampling portion in order to preserve spatial information. It is called U-Net because the resulting architecture is depicted in a U-shaped sequence of operations in a diagram of the model.

Pix2Pix [44] is another such image segmentation architecture based off of U-Net which is applicable across many different multi-class classification tasks including image segmentation. The decoder for this model which carries out the up-sampling layers is freely available in TensorFlow examples and can be combined with pre-trained encoders such as MobileNetV2 introduced previously by using skip connections from intermediate layers.

1.3.4 Performance metrics, validation and uncertainty

A number of performance metrics exist to characterize the data retrieved upon application of a prediction by considering errors on positive or negative examples. If the prediction is positive and it is correct then it is called a true positive T_p, if it's incorrect it is called a false positive F_p, and similarly for negative examples (T_n and F_n).

The precision is defined in terms of true and false positives as

$$\text{PRECISION} = \frac{T_p}{T_p + F_p}$$

and is used to quantify how many retrieved positive elements are correct out of all positive elements. For example, this would quantify the fraction of male fruit

flies retrieved in a collection basket of true males and false males (aka females). A high precision implies that the collection of positive retrieved elements has low contamination with false positives. This essentially calculates the purity of the positive collection basket by estimating the probability that a male in the male basket is actually male.

The recall on the other hand quantifies how many correct positive elements are retrieved out of all possible positive elements that could have been retrieved. Thus the recall is defined as

$$\text{RECALL} = \frac{T_p}{T_p + F_n}$$

and it would for example quantify how many male fruit flies were retrieved out of the total available male fruit flies. A low recall implies that many males fruit flies are incorrectly classified as female and put into the female basket. This is also called the true positive rate (TPR).

The false positive rate (FPR) is called the specificity, which quantifies how many negative elements are correctly classified as negative. The specificity is defined as

$$\text{SPECIFICITY} = \frac{T_n}{T_n + F_p}$$

and a high specificity means little contamination of the positive basket with false positive examples. For example, a high specificity means that there are few female fruit flies put into the male basket.

If the data set is balanced one can simply compare the number of false negatives and false positives to get an idea of how the algorithm is performing. One advantage of these performance metrics is that if the data set is not balanced then they still give information in a format that may carry over to operational settings,

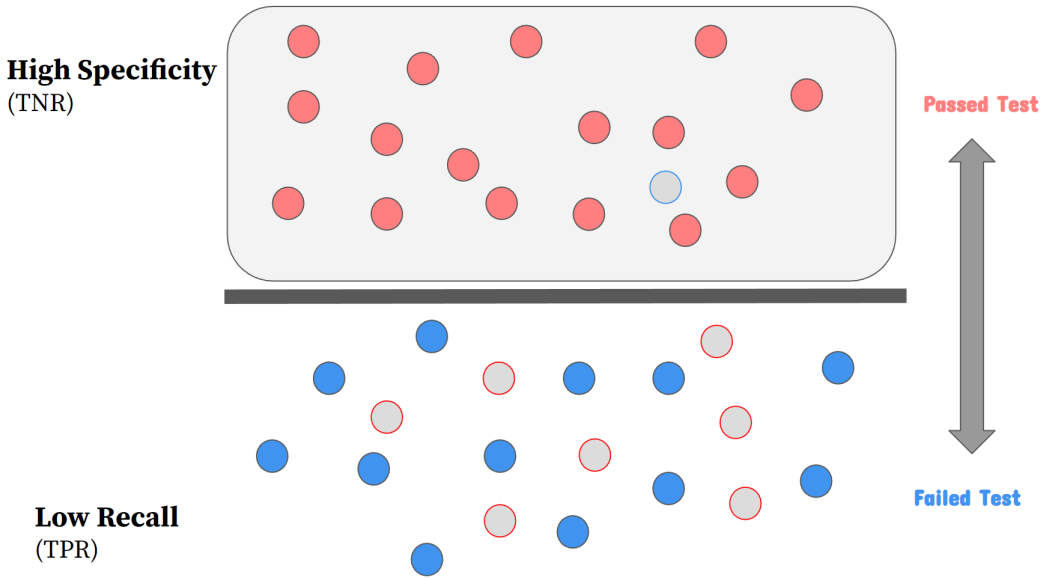

Figure 1.9: This is an example of high specificity and low recall. A classifier with low recall loses desired class instances by accidentally putting them into the failed test bin. For SWD recall is the number of males in the passed-bin out of the total males available. Specificity is the number of females in the failed-bin out of the total females available.

provided the distribution in the operational setting is close to the experimental setting.

It is essential that these metrics are calculated on validation data, which has not been used for training, for otherwise the model would be biased to correctly classify elements which may not have been correct if they were not included in the training data, as will be the case for elements observed in an operational setting. If the number of validation samples used to make these comparisons is small then further steps can be taken to estimate performance in the general setting. A common method that attempts to extrapolate results of training a model to the operational setting without discarding the validation data is called k-fold cross validation. This method splits the available data set into a number of folds k, where $\frac{1}{k}$ are used as validation data for extrapolation to the operational setting, and $\frac{k-1}{k}$ of the remain-

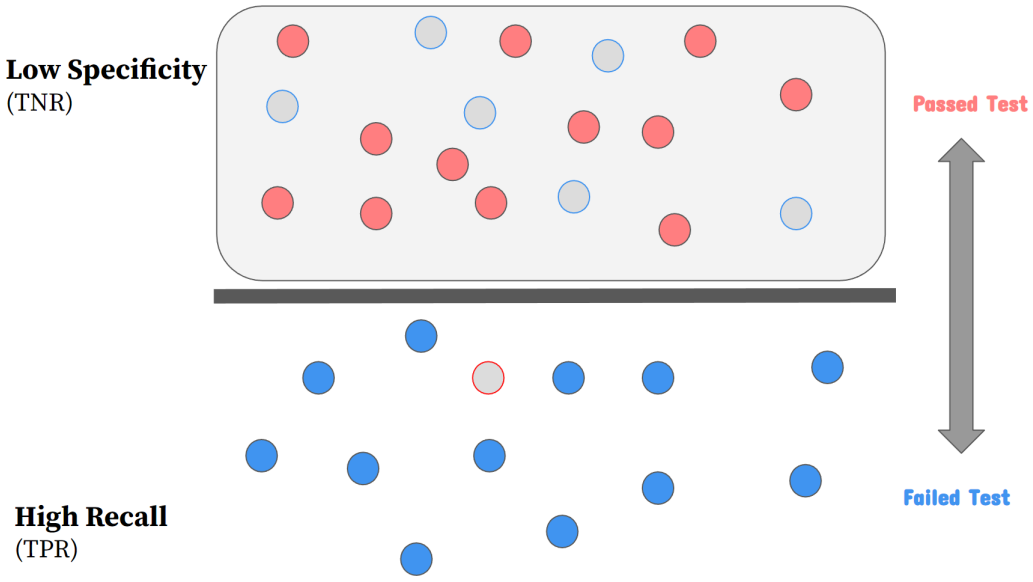

Figure 1.10: This is an example of low specificity and high recall. A classifier with low specificity loses negative examples to the positive class (with false positives).

ing folds are used for training data. This process is repeated again by using one of the other folds as validation and placing the already-used fold back into the training data set. Averaging over the performance results of k different models each trained separately and gives an idea of the standard deviation of the results. More sophisticated methods exist to utilize even more of the data, but typically a splitting of the data into for example $k = 5$ folds is acceptable. While these performance metrics and validation methods are useful for estimating the effectiveness of the model on future unseen data and for informing the relevancy of the results, further analysis may be carried out to investigate how the model sees individual examples.

If the classification algorithm outputs a probability for the prediction (or if it can otherwise be extracted) then the entropy of a prediction can be useful for quantifying how confident a model's predictions are. The Shannon entropy is defined

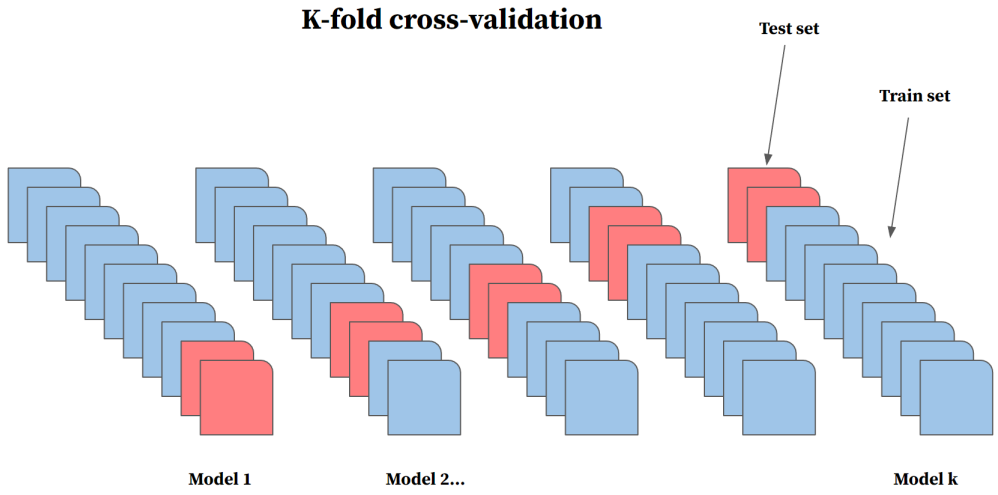

Figure 1.11: This is a depiction of k-fold cross-evaluation where k=5 and the evaluation data is split into 5 different sets with associated training data. One model is trained for each set and the metric of interest for each training run is averaged over all of them.

in terms of a probability distribution $p(x)$ for outcome variables x as

$$\mathcal{E} = -\sum_{x} p(x) \ln p(x).$$

This is useful for quantifying the confidence of a prediction from a model; if the probability distribution over prediction variables is close to uniform, then the entropy is higher and the confidence of the prediction is lower. If on the other hand the probability of a prediction is close to 1 for one outcome and close to zero for the others, then the entropy is close to zero and the confidence of the prediction is high. This can give information about which data points may have been incorrectly classified with high confidence. If such a case happens then it is possible that the model will not generalize well to similar data points. Higher entropy examples may also give some information about the prediction, for example if it is a difficult sample for the model or there were many data points in the data set that looked similar to each other but belonged to separate classes. The entropy may be customized

to emphasize particular prediction variables by taking only a partial sum over the relevant variables. Further care may be needed to understand the limitations of this type of usage.

Some classification algorithms do not estimate a probability directly of a binary classification, but rather output a decision with no further information. In this case an algorithm called LIME [45] may be used. LIME stands for local interpretable model agnostic explanations and was an attempt to understand predictions output by a model by sampling data with perturbed patches of features (for example pixels in an input image) to produce synthetic data and by measuring how unfaithful the synthetic data is at approximating the decision of the model on the original data. This gives a way to output a representation of the features that were important in classifying a particular input sample. A canonical example presented is one in which a husky is classified as a wolf, and LIME demonstrates that the snow in the background picture was key to making the identification of the wolf. This is of course spurious, as huskies are also observable in snowy environments; however it may be an indication that the data set was biased. The disadvantages of this algorithm are that it takes potentially a long time to produce predictions from slow models and that the effectiveness of the output of the explanation may be dependent on the domain and may not give information as specific as what the inner workings of the model are capable of producing with some modification.

Chapter 2

Classification of Drosophila suzukii

2.1 Overview

Drosophila suzukii is an invasive species of fruit fly responsible for causing concerning amounts of damage to the global supply of soft and stone fruits. Some estimates indicate that there is potential for over $700M worth of crop damage due to SWD every year in the United States alone [7, 8, 9]. Sterile insect technique (SIT) has been identified as a viable means for control of this pest[24]. One of the major challenges identified for the deployment of SIT is the automation of sex separation in the process of collecting enough sterile insects [25]. In order to separate male from female specimen a classification algorithm is needed to automatically identify the sex of the insect in an image taken by the processing system.

A review of studies of image capture and classification of insects shows a recent trend toward applying deep learning methods compared to shallow learning methods for automatic identification of insects [32]. One recent work demonstrated that deep features in algorithms like VGG16 and VGG19 pretrained on ImageNet in combination with SVM and other classifiers can achieve excellent accuracy for insect

species identification from high resolution pictures [38]. On the other hand, systems based off of the YOLO algorithms for localization and identification of insect species and gender have also shown success, especially when the number of wild-caught biological samples is large enough for a sufficient training data set [34]. Approaches like these typically augment the data sets to produce an order of magnitude or more the amount of data compared the original data sets.

This work investigates whether pre-trained networks are useful for classification of sex on a newly collected data set of *Drosophila* suzukii images taken from microscope slides. The recent CLIP algorithm and its associated models have shown great promise in solving many classification tasks [2]. Deep features from six different CLIP models are compared to three networks pre-trained on ImageNet: VGG16, VGG19 and MobileNetV2. Binary classification tasks are probed with logistic regression and SVM classifiers. Impressive accuracy is achieved even before data augmentation and for microscope slides captured at lower resolutions than tested for VGG16 and VGG19 in [38]. The main contributions of this work are the collection of a *Drosophila* suzukii data set categorized by male and female and a demonstration that models such as CLIP trained on massive amounts of internet data can be successful at classification of biological data and achieve higher accuracy than well-established models such as VGG16 and VGG19 and MobileNetV2.

2.2 A data set of SWD images

Experimental settings

The experimental setup will be discussed in this section. Details about the data set, its collection and modification will be presented with the image processing

Scope Magnification	# of examples
27x	429
17x	271

Table 2.1: Shown is the number of samples used for training data from each approximate scope magnification. The 17x data set utilizes the clone tool from GIMP to remove specimen that are overlapping and may have deleterious effects on certain aspects of learning, while preserving the important features for identifying between male and female.

and machine learning strategies used in this work.

2.2.0.1 Introduction to the data set

The data set used for image classification and segmentation in this work is comprised of 676 Drosophila suzukii images, evenly split between categories of 338 males and 338 females.

2.2.0.2 Collection of the data set

The Drosophila suzukii specimens' images studied in this experiment were collected from the Santa Cruz region by researchers based out of UC Santa Cruz. The specimens were identified as positive for Drosophila suzukii before being housed at the Sinsheimer labs for Biological Sciences on campus at UC Santa Cruz, where they were imaged with a Dino-eye lense at magnifications of approximately 17x and 27x, which generated microscope images of pixel-size 2592×1944. Because of the limited availability of a large number of specimens, some flies are imaged under different position and lighting conditions to create additional data.

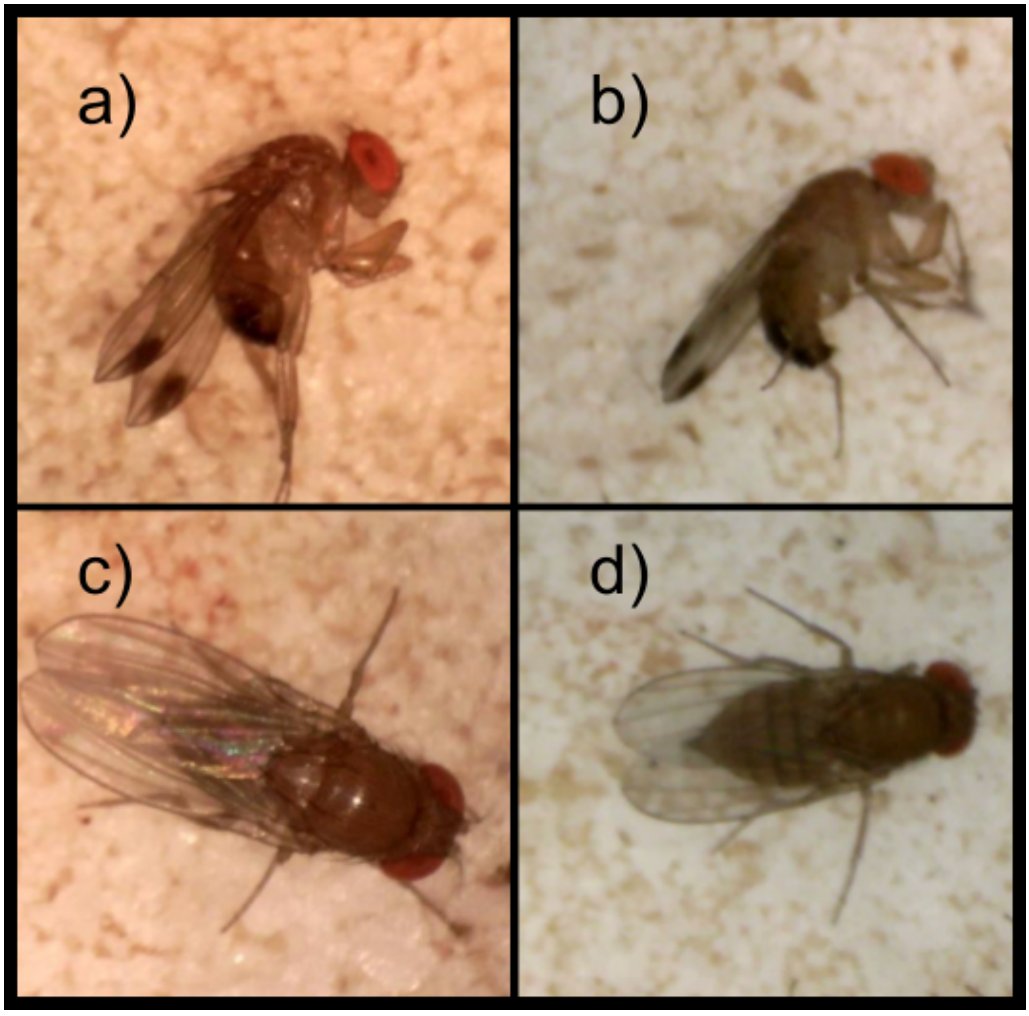

Figure 2.1: Four example images of SWD taken at different magnifications: (a) image of a male *D. suzukii* at 27x magnification, (b) image of a male *D. suzukii* at 17x magnification, (c) image of a female *D. suzukii* taken at 27x magnifcation, (d) image of a female *D. suzukii* taken at 17x magnification.

2.2.0.3 Pre-processing the images

In order to build a data set consumable by a wide variety of machine learning pipelines screen-captures of the microscope slides were taken such that each screenshot fit each specimen into a 512×512 pixel image. The images were then classified and labeled with "m" for male or "f" for female within each filename before the file extension. The microscope images taken at the lowest magnification (17x) contained fruit flies of the opposite sex, which may not be problematic for some generalized object detection algorithms, but were an unnecessary complication for the machine learning approach presented in this work where the industry application would focus on single specimens on an assembly line. The clonetool provided by the GNU Image Manipulation Program (GIMP) was used to effectively remove specimens which contaminated images centered on a main target specimen. At higher resolutions these images where overlapping parts of flies, such as the legs, may have a small effect on classification efficacy (e.g. of the sex combs on male specimens' legs), but at the resolutions used within this work they are not readily visible. Images where the main body, wing or other important visible features such as the serrated ovipositor are obscured by nearby flies are discarded. In addition to the binary classification tag for the sex of the specimen, segmentation images delineating the body of each fly from the background were hand-drawn with a mouse.

2.3 Binary classification with pre-trained networks

The pre-trained models and probe classifiers use for classification of the data set are described in this section.

2.3.1 Models pre-trained on ImageNet

2.3.1.1 VGG16 and VGG19

The VGG net models, VGG16 and VGG19, with 16 and 19 layers of small 3x3 ConvNet models come from submissions to the ImageNet Challenge 2014 competition where they achieved first and second prize in localization and classification respectively [39]. VGG16 and VGG19 are relatively slow models with approximately 138 million and 143 million parameters respectively, though it was found that about 100 million could removed from the fully connected layer without losing performance. The goal of the VGG models was to increase classification accuracy by increasing the depth of ConvNets.

VGG net has an input size of 224x224 RGB images, and passes its input through either 16 or 19 3x3 convolutional layers of stride-1, with five layers of max pooling each downsampling between the convolutional layers. Behind this there are three fully-connected layers followed by a softmax layer.

2.3.1.2 MobileNetV2

The MobileNet series of neural networks are developed to improve efficiency so that machine learning models can be used on mobile devices. MobileNetV2 was designed to be more efficient by making convolutions from different channels separable, which reduces computational overload while performance dips only slightly [1]. It consists of a bottleneck of convolutional layers similar to how the VGG16/19 models operate, but the central layers are formed with units of depthwise convolutions and the layers in the center are repeated to make it deeper.

2.3.2 CLIP models pre-trained on internet data

2.3.2.1 ViT-B/16 and ViT-B/32

The visual transformers for CLIP closely follow the implementation from [3], which splits an image up into patches and combines patch embeddings with positional embeddings and an additional classification token, before being put into a the transformer encoding, which outputs into a MLP layer for classification. The main difference between the original model and the visual transformers used in CLIP is that there is an extra batch normalization layer between the patch-position embedding and the encoder input.

2.3.2.2 RN50x4 and RN50x16

The RN50x4 and RN50x16 models are ResNets trained in the style of EfficientNet from [6], which scales the depth and width (e.g. layers and channels) of ResNet's parameters with a fixed ratio.

2.3.2.3 RN50 and RN101

The RN50 and RN101 models are also ResNets but with improvements from a the ResNet-D module from [46] which improves performance in the down-sampling blocks. The modifications for CLIP also implement [5] which uses BlurPooling to fix aliasing for down sampling between convolutional layers. The last modification is an attention pooling mechanism that the last layer ends with, where the query is conditioned on the global average-pooled representation of the image.

2.3.3 Probe classifiers

Four classifiers were used to test the effectiveness of image encoding from each pre-trained network: a support vector machine (SVM) with a linear kernel, another implementation of a linear SVM optimised with stochastic gradient descent (SGD), an SVM trained with a polynomial kernel of degree 3, and a logistic regression classifier. Before input into these classifiers the outputs of the embeddings are shifted by mean and scaled by the variance of their values during the optimization process.

These methods are implemented using Python 3.7.6, TensorFlow2, PyTorch 1.7.1 and Scikit-learn 1.0.2. All training, testing and development of these pipelines were run within Google Colab sessions, which have variable GPU types, amounts of RAM and processing speed.

2.4 Results

The results of the experiments with pre-trained models from CLIP data and ImageNet data adapted to learn the sex of SWD are presented in this section.

Accuracy				
Pretrained model	SGD	LinSVC	PolySVC	LogReg
ViT-B/32	0.9230 ±0.0148	0.9306 ±0.0187	0.9261 ±0.0278	0.9365 ±0.0280
ViT-B/16	0.9571 ±0.01272	0.9571 ±0.0144	0.9512 ±0.0087	0.9674 ±0.0103
RN50x16	0.9230 ±0.0266	0.9262 ±0.0239	0.9202 ±0.0131	0.9203 ±0.0253
RN50x4	0.9274 ±0.0262	0.9231 ±0.0182	0.9333 ±0.0173	0.9261 ±0.0191
RN101	0.9201 ±0.0259	0.8964 ±0.0126	0.9349 ±0.0146	0.9113 ±0.0146
RN50	0.9348 ±0.0256	0.9290 ±0.0108	0.9247 ±0.0222	0.9336 ±0.0256
VGG16	0.8890 ±0.0107	0.9142 ±0.0211	0.4794 ±0.0239	0.9172 ±0.0252
VGG19	0.8817 ±0.0278	0.9172 ±0.0169	0.4836 ±0.0464	0.9096 ±0.0337
MobileNetV2	0.8682 ±0.0205	0.8906 ±0.0258	0.5360 ±0.0684	0.9007 ±0.0297

Table 2.2: A table showing accuracy for binary classification with different pretrained models. To augment the data, the images were flipped vertically and horizontally once, multiplies the data by 3x. The model ViT-B/16 performed the best with a logistic regression classifier but also performed well with all SVC classifers.

Specificity				
Pretrained model	SGD	LinSVC	PolySVC	LogReg
ViT-B/32	0.9258 ±0.0204	0.9341 ±0.0302	0.9137 ±0.0306	0.9415 ±0.0204
ViT-B/16	0.9447 ±0.0182	0.9480 ±0.0221	0.9400 ±0.0259	0.9693 ±0.0105
RN50x16	0.9020 ±0.0430	0.9045 ±0.0367	0.8785 ±0.0331	0.9175 ±0.0272
RN50x4	0.9188 ±0.0478	0.9216 ±0.0250	0.9083 ±0.0288	0.9447 ±0.0332
RN101	0.9009 ±0.0460	0.8840 ±0.0278	0.8971 ±0.0188	0.8983 ±0.0219
RN50	0.9262 ±0.0434	0.9314 ±0.0256	0.8935 ±0.0421	0.9182 ±0.0429
VGG16	0.9202 ±0.0450	0.9181 ±0.0405	0.0358 ±0.0185	0.9135 ±0.0291
VGG19	0.8697 ±0.0777	0.9139 ±0.0113	0.0422 ±0.0153	0.9154 ±0.0400
MobileNetV2	0.8245 ±0.0259	0.9059 ±0.0341	0.1529 ±0.0965	0.9164 ±0.0413

Table 2.3: A table showing specificity for binary classification with different pretrained models. The model with highest specificity was the ViT-N/16 version from CLIP. A classification algorithm with high specificity experiences few false positives. A high specificity in classifying positive samples (males) means that the males selected for SIT will have a a low amount of false positive females in the collection of males.

Recall				
Pretrained model	SGD	LinSVC	PolySVC	LogReg
ViT-B/32	0.9217 ±0.0296	0.9242 ±0.0336	0.9398 ±0.0265	0.9294 ±0.0523
ViT-B/16	0.9717 ±0.0180	0.9683 ±0.0146	0.9660 ±0.0231	0.9635 ±0.0336
RN50x16	0.9480 ±0.0256	0.9534 ±0.0211	0.9687 ±0.0170	0.9245 ±0.0289
RN50x4	0.9409 ±0.0414	0.9244 ±0.0317	0.9619 ±0.0246	0.9135 ±0.0545
RN101	0.9438 ±0.0401	0.9084 ±0.0166	0.9779 ±0.0167	0.9248 ±0.0133
RN50	0.9442 ±0.0250	0.9272 ±0.0151	0.9618 ±0.0077	0.9507 ±0.0172
VGG16	0.8504 ±0.0445	0.9061 ±0.0372	0.9840 ±0.0108	0.9245 ±0.0479
VGG19	0.8937 ±0.0634	0.9207 ±0.0263	0.9869 ±0.0133	0.8977 ±0.0591
MobileNetV2	0.9174 ±0.0197	0.8736 ±0.0450	0.9785 ±0.0151	0.8848 ±0.0257

Table 2.4: A table showing recall for binary classification with different pre-trained models with augmentation. The highest recall shown was using the PolySVC classifier which failed to converge for VGG and MobileNetV2 models, leading to all predictions that were all female. The best results consistent with high recall and high specificity are again for ViT-B/16. While the ResNet-based clip models also had high recall, their specificity was a few points lower than the ViT models.

Precision				
Pretrained model	SGD	LinSVC	PolySVC	LogReg
ViT-B/32	0.9142 ±0.0288	0.9293 ±0.0221	0.9058 ±0.0302	0.9334 ±0.0241
ViT-B/16	0.9384 ±0.0234	0.9410 ±0.0276	0.9321 ±0.0327	0.9648 ±0.0121
RN50x16	0.8970 ±0.0411	0.8956 ±0.0434	0.8754 ±0.0315	0.9058 ±0.0378
RN50x4	0.9092 ±0.0590	0.9135 ±0.0204	0.9015 ±0.0312	0.9300 ±0.0522
RN101	0.8940 ±0.0477	0.8750 ±0.0159	0.8928 ±0.0200	0.8901 ±0.0133
RN50	0.9224 ±0.0391	0.9211 ±0.0306	0.8860 ±0.0506	0.9146 ±0.0403
VGG16	0.9110 ±0.0431	0.9105 ±0.0346	0.4726 ±0.0237	0.9048 ±0.0279
VGG19	0.8629 ±0.0827	0.9029 ±0.0200	0.4749 ±0.0442	0.9052 ±0.0444
MobileNetV2	0.8216 ±0.0214	0.8892 ±0.0422	0.5050 ±0.0566	0.9038 ±0.0491

Table 2.5: A table showing precision for binary classification with different pretrained models with augmentation. Precision is proportional to the fraction of true positive (males) detected out of the total number of predicted males.

2.4.0.1 Augmentation tests

A number of augmentations were tested, including inversions, the RandAug function, image rotations, color threshold alterations. Comments on a few of these are included in the supplementary sections relevant to this chapter. Due to the small size of the data set and close resemblance of the training data to the test data, large amounts of augmentations are avoided. Some of the augmentation strategies could be further explored (e.g. with a smaller magnitude or more precise alteration to the image in order to better match the test data distribution). The simplest augmentation strategy of using horizontal and vertical flips turned out to be the best among the strategies tested.

2.5 Discussion

The accuracy for each model is presented in table 2.4. Results indicate that the CLIP models are consistently better at learning from the SWD data set than the models trained on ImageNet. While the results from VGG16/19 and MobileNetV2 are still impressive, almost reaching above 90% accuracy, the CLIP models uniformly perform better when trained on the same data set with the same probe classifiers. This is not unexpected but still somewhat surprising because in [2] the authors indicate that CLIP is week at zero-shot tasks that are highly specialized or complex such as identifying satellite images, lymph node tumor detection, counting objects in synthetic scenes and more. Preliminary tests indicate that CLIP has a difficult time identifying gender of SWD in zero-shot settings while using only natural language prompts compared to images using the text-encoder included with the CLIP models. Nonetheless when the image encoding from CLIP is adapted with LogReg or SVC

algorithms, accuracy eclipses 92 % with all CLIP models tested and is observed as high as 96.7% with ViT-B/16 + LogReg in this experiment. All of the models showed impressive specificity and recall except for the PolySVC probe classifier, which had some trouble fitting to VGG and MobileNetV2 embeddings and did not converge at the tolerances tested. These impressive results could be further improved with more specialized augmentation, iteration and extension of the classifier models operating on the output of image encoding from the pre-trained models, model combination and with larger amounts of data specialized for the rearing or laboratory settings.

Conrastive pre-training with natural language is further corroborated by these results which show that the resulting image encoding from CLIP models outperform models used widely in for bench-marking and industry applications. Extending applications of CLIP's image encoding are not limited to this task of identifying sex of SWD, as the same embeddings can be calculated once and used in conjunction either with another adapted classifier or the text encoding that is also provided by CLIP for zero-shot settings, which may be useful for example at identifying if there are other objects left in the camera scene while identifying the sex of SWD (such as another type of insect or other obstructive smudges or objects). While models of the CLIP variety have immediate applications where it may be useful, it is likely that zero-shot models like CLIP and its text-based ancestors like GPT-3 will continue to improve to the point where even the specialized and complex tasks will be solvable with minimal (or no) description of the task. This work serves as a demonstration of the wide applicability of available machine learning models and also as a snapshot of AI research which is rapidly moving towards a future where extensive testing and adapting of models by humans will automated by large models trained on massive public data sets.

In the near term, these results show that the main objective to classify the sex of SWD is achievable for a industrial or laboratory setting with pre-trained networks and with small data sets of varying magnification. Further work could still be done to evaluate what data and how much is needed to extend the effectiveness of these models to settings which will be useful out in the field on the farm or in other natural settings which host SWD. Trapping systems which are useful for monitoring presence of SWD could be automated to provide evidence of an outbreak on a farm without need for tedious observation by specialized personnel who have been trained to identify SWD among other species which may be caught in a trap. It would potentially be helpful to identify different out-of-disribution data which would diversify the robustness among different tasks outside of the research settings explored in this experiment.

Chapter 3

Segmentation of Drosophila suzukii

3.1 Overview

Spotted wing drosophila (SWD) is an invasive species of fruit fly which has in the last decades spread from east Asia across the Pacific Ocean to the mainland of the United States, where it has proliferated across the entire country. While in the previous chapter an effective set of algorithms was used to demonstrate efficacy in binary classification of male and female SWD for applications to control the spread of this pest, questions about the viability of the data for use-cases beyond the laboratory setting would inform further data collection efforts aimed at these use-cases, for example under man-made plastic backgrounds which may be used in traps, or for other scientific purposes such as observation of SWD under natural conditions and at magnifications and lighting conditions not explored in the original work. While analysis of binary classification strategies gives a strong idea of the performance of a particular model on a data set, further exploration of the explanation or reasoning for a model's prediction is sometimes desired for a training data set. One such machine learning strategy that gives insight into the training data is image segmen-

tation, where the individual pixels belonging to the object being identified in the image are also output by the model. The only disadvantage of segmentation is the extra effort it takes to produce a training data set; though many studies have set out to automate this process too by reducing the amount of data needed to produce segmented versions of the data set.

The objective of this work is to investigate the overlap provided by this laboratory segmented data set on out-of-distribution (OOD) and to use image segmentation as an alternative means to explore the morphology of SWD. In addition to the use of image segmentation to analyze both in-distribution data and OOD data, specialized data augmentation obtained from the segmentation is explored as a strategy to improve robustness in classification results. Image encoding from MobileNetV2 is used in conjunction with Pix2Pix, a modified version of U-NET for biological image segmentation, to classify between male and female specimen of SWD. The confidence of the model's prediction is defined in terms of entropy. Evaluation of the machine confidence are presented and analyzed to show three main results. (1) The introduction of specialized augmentation shows increased confidence from the model under evaluation of various OOD challenges for SWD when aggregated over predictions from regions of certain pixel classification, (2) that closer inspection of the pixel-wise machine confidence can yield the morphological features responsible for positive or negative classification of SWD, and (3) that this segmentation model trained under laboratory conditions retains slight ability to give insights on OOD data not included during training, particularly when the resolution is high enough to show details on morphological features.

3.2 Results

This section will first describe the creation of a small OOD data set intended to test the limits of models trained on the laboratory SWD data set. The next subsection will outline the creation of a segmentation data set based off of the laboratory data set. Following that, the pre-trained model used for image encoding and its decoder used for segmentation will be described before the augmentation strategy is covered in the next section. Finally, quantitative and visual results of machine confidence from experiments of three different resolutions are presented.

3.2.1 An out-of-distribution data set

A small data set of 76 SWD examples (38 male and 38 female) are taken with a Sandmarc macrolense at 10x magnification from an an iPhone 13 mini camera. All images are taken at a local wild raspberry patch located in Santa Cruz which harbor wild SWD. A SWD attractant was concocted and traps were created from used water bottles following the advice from [52] closely to observe the presence of SWD at the raspberry patch. While the water bottle trap was capable of collecting SWD its design was not optimized for taking photographs; even though it was transparent moisture from the attractant reduced visibility and therefore all traps were discarded. The attractant was fortunately very potent and attracted many SWD which allowed for photography of the insect. Photographs were taken in the late morning with natural lighting conditions and in the early evening with a supplemental LED light.

Examples from the data set are shown in figure 3.1. The first two images are males, one on a plastic lid used as a drop-in substitute for plastic that might

be used on an outdoor SWD surveillance device, and the other male SWD is on a leaf. The other two images are of females, one on a different leaf and the other on a raspberry. The pictures in the data set were cropped in order to roughly fit the flies onto the image size, which is 512x512, like the original laboratory set.

This data set is pooled for analysis to the additional data leftover from the original laboratory data set. These additional laboratory images comprised of SWD images which were unused for training because they were more difficult due to important features being obscured by the lighting and positioning of flies in the image, or because part of the fly was severely cut off in the image; however they retain an important role in analysis to gauge what kind of errors can arise due to non-ideal conditions when the fly is not completely visible. This data set is referred to in figures as the "leftover set". A secondary small data set of laboratory examples also unused as training data whose sexes were difficult for this author to identify were analyzed separately; this set is referred to as the "challenge" or "difficult" set in figures.

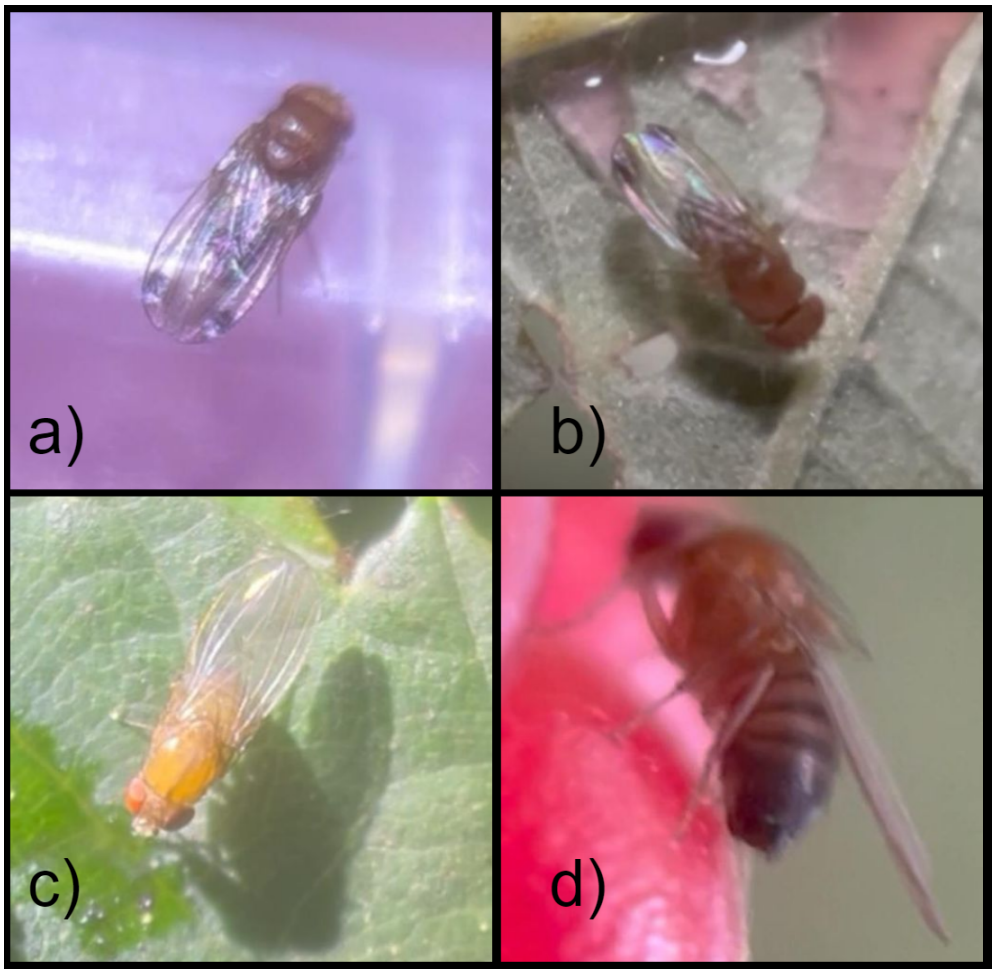

Figure 3.1: An out-of-distribution (OOD) data set is collected for the purposes of evaluating the robustness of segmentation results on data taken outside of the laboratory, which is referred to as the OOD data set. This data is used to evaluate segmentation results for unusual cases far away from the intended use-case. Shown are some example images from the data set of SWD under natural and artificial lighting conditions: (a) a male on a man-made surface, (b) a male on a leaf, (c) a female on a leaf and (d) a close-up of a female on a raspberry.

3.2.2 Segmentation training data

The segmentation of the laboratory data set was carried out by mouse and hand in Microsoft Paint by erasing the relevant portion of the image and saving it. Software was written to convert the color image into a mask of integers corresponding the associated sex and pixel type (either edge, body or background pixels) of the SWD. The edge pixels were generated automatically by translating the zeroed out body mask along multiple directions by 3 pixels and taking the non-zero difference between pixels of the central original mask and each translated mask and adding them together. Once the edge mask is calculated it is labeled with the appropriate multi-class label. These included background pixels, female body pixels, female edge pixels, male body pixels and male edge pixels.

This technique was effective enough at generating masks for the training data set. For the OOD data set where color variation includes perfect white pixels due to bright artificial lighting from the LEDs there could be some artifacts brought into the segmented masks. While they are not used for training any of the models in this work, an improved version would use an industry software to create the mask data set.

An example result from the segmentation data set is shown in figure 3.2 for a female. Many fine details are captured from the segmentation work, including traces over the legs of fly. The algorithm must be able to distinguish between the background and the transparent wing of the fly, which could be anticipated as a more challenging task than identifying the opaque body of the fly. The wings and the serrated ovipositor at the bottom of the fly are some of the most important features needed to identify the sex of the fly; however at high resolution the sex combs may

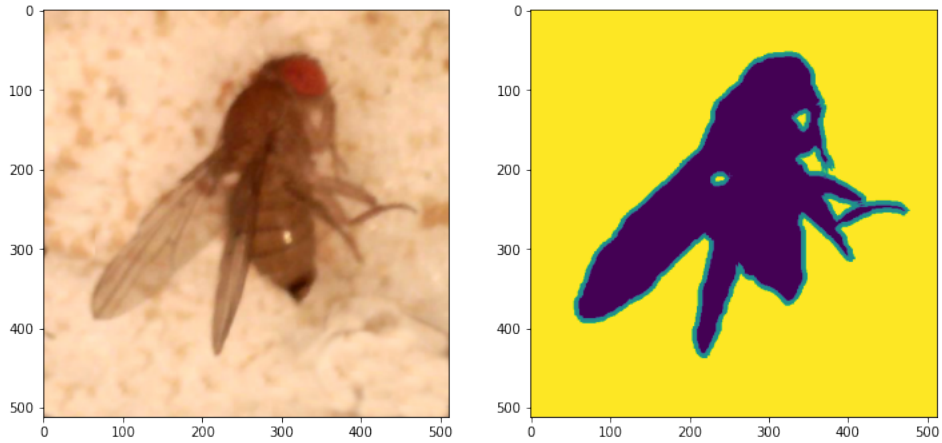

Figure 3.2: Shown (left panel) is a specimen from the original laboratory data set and (right panel) the corresponding segmented example from data set of segmented training labels. For this depiction of the segmented example, the background is yellow, the body of the fly is dark blue and the edge pixels are green.

be visible on the legs of the male, though the accuracy of hand segmentations may impede the efficacy of leg features if they are even visible at the resolutions tested.

3.2.3 Augmentation strategies used during segmentation training

Multiple sources of augmentation that were used were random alterations to the images that included changing attributes such as the image quality, brightness, hue and saturation, as well as application of Gaussian blurs. A specialized augmentation tested for the purposes of this experiment was to be able to cut and paste rotated flies on a set of background slides to see if that would help expand the data set in an effective way. An example of this rotation is shown in figure 3.3. The general observation with the augmentation is that it must be limited in scope to make the newly minted augmented training data somewhat similar to what one expects to see on the validation data. Most of the augmentation used in this work is not up to the task of altering the image in a way that it is similar to the difficult examples or the out-of-distribution data. In fact in many instances the application

Figure 3.3: One of the desired augmentation strategies was to rotate the fly on the image itself to remove spatial correlations between parts of the fly and parts of the background. Shown here is the fly rotated and overwritten onto the same background, though during augmentation a random background was chosen to remove correlations between specific backgrounds and their corresponding fly.

of the augmentation reduces the accuracy on all of the validation data regardless of its quality; however it does affect the confidence of the model prediction. With more careful administering of these augmentations it is likely more improvement can be made on the final binary classification accuracy of the model.

3.2.4 Pre-trained image encoder with trainable decoder

Choosing a pre-trained model for development of a segmentation was done so that the compute time would not be as expensive with minimal concern towards accuracy. MobileNetV2 was therefore a choice that carries overlap with the previous chapter's results while being speedy and maintaining respectable accuracy. The segmentation model that uses these embeddings is a modified version of U-Net called Pix2Pix and is frequently used as a demo model by the community. Future work can explore the likely performance gains one will obtain by using the CLIP and other models in conjunction with other decoders.

3.2.5 Notable SWD features in machine confidence

In this work a customized application of entropy is used to gauge the confidence of the model's prediction and the effects of augmentation on this confidence. At the pixel level, a separate application of entropy is used to visualize the confidence of the model's prediction under different circumstances of data input.

3.2.5.1 Pixel-wise entropy and selected comparisons

To visualize the confidence of the model's output, the entropy can be calculated over the five classification variables for each output pixel. Pieces of SWD morphology and variations between sexes can then be identified within segmentation and entropy visualizations.

Throughout this chapter images of pixel-wise entropy and partial entropy have a color scheme for each image that goes from zero to one, with dark blue being closer to zero and bright lime being closer to one, where the data is normalized such that the minimum entropy is zero and the maximum entropy is one. The same applies to visualizations of partial entropy (for example in comparing edge-pixel probabilities only).

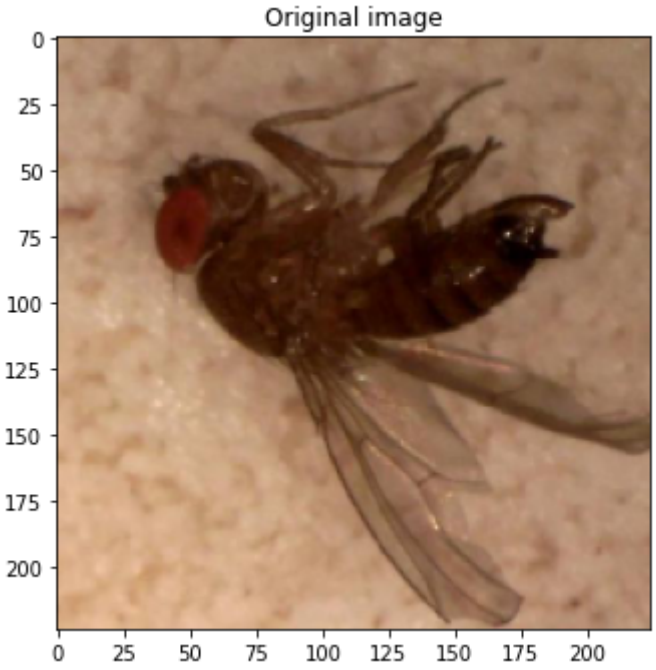

Figure 3.4: A size 224x224 female SWD from the training data set.

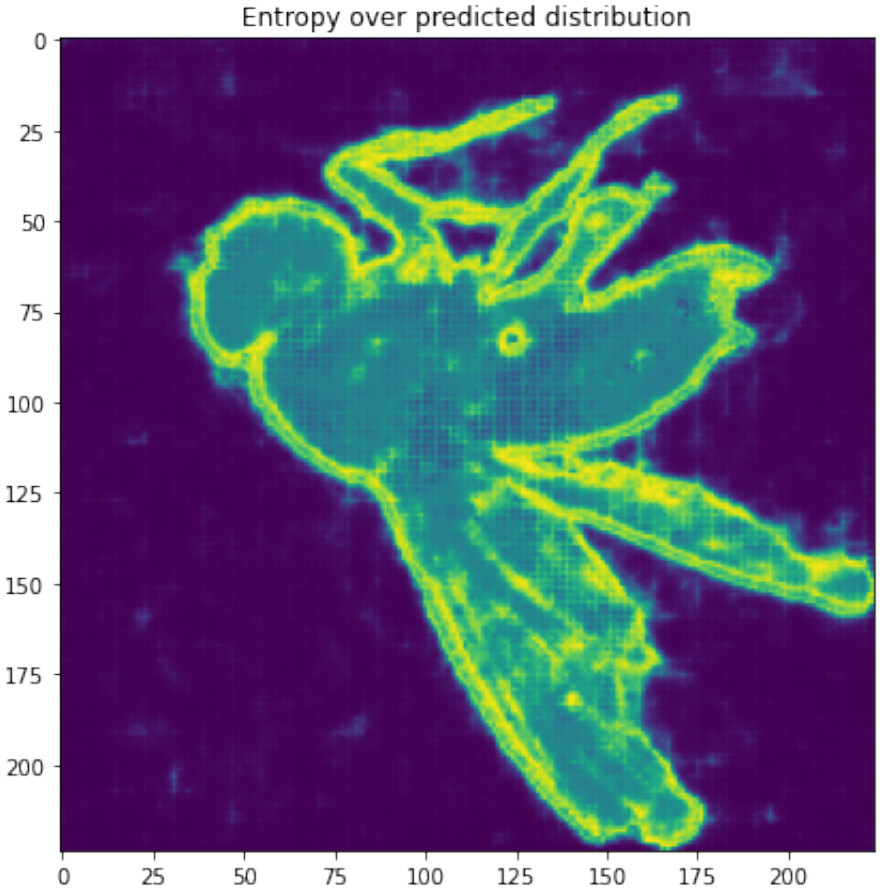

Figure 3.5: Visualization of the pixel-wise entropy of a 224x224 female SWD from the training data set. The watermark from the 3-pixel translations to create the edge-classified pixels is clearly visible around the outline of the image; these pixels are likely the highest entropy because there is most variation among the and drawings and the edges are some of the highest contrast pixels used to differentiate body from background. The transparent female wings have higher entropy along the creases within the wing as well.

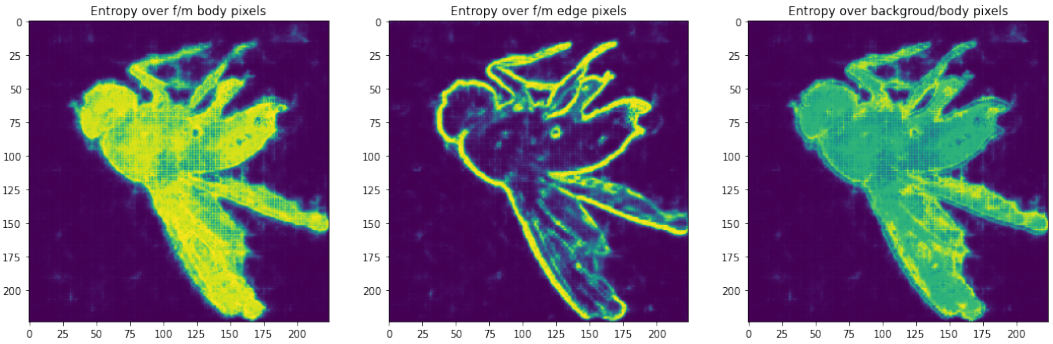

Figure 3.6: Visualization of the normalized partial-sum of pixel-wise entropy from selected groupings of pixel classification of a female SWD from the training data set. (Left panel) The partial entropy over pixels classified parts of the male or female body. For this example the lower entropy points are along the bottom part of the wing and towards the center of the body. (Middle panel) Since this example is from the training set over-fitting was likely occurring on the details of the features inside the body pixels of the fly and have strong contrast consistent with the augmentation and edge-creation procedure. Nonetheless it is natural to expect parts of the wing to have high entropy since it is transparent. (Right panel) The partial entropy between the body and background pixels.

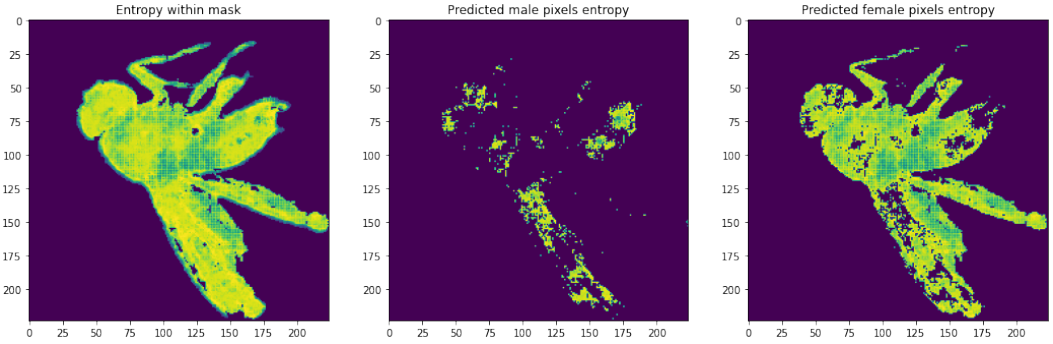

Figure 3.7: Visualization of the pixel-wise entropy of a female SWD from the training data set cut into portions that were (left panel) classified as body pixels, (middle panel) classified as positive for male-body pixels and (right panel) classified as female-body pixels. For this training set example it is very clear that the prediction is going to be female, since there are more female pixels. When you only keep entropy to a threshold of $\varepsilon \leq 0.3$, only 1 pixel of male-body classification remains and there are 7 pixels of female body classification. Whereas before the threshold there was 1944 male pixels and 10972 female pixels. The ratio between female-body to male-body pixels widened from 5.64 to 7.0 when taking the lower entropy pixels. The same happens for edge pixels but is for obvious reasons less reliable than for body pixels since there are fewer edge pixels to train on than body pixels. The ratio in this example for edge-specific pixels goes from about 1.6 to 2.1 after the entropy threshold.

3.2.5.2 Identifying features of SWD with pixel-wise entropy

In this section an example segmentation for a male from the validation set will be presented and described in the context of its morphological relevance to the classification algorithm, namely its spotted wing and darkened backside.

One result demonstrating this is shown in figure 3.8: a male from the validation set. The classifier identifies most of the fly as male. To check which portions of the image are most strongly identified as male, the partial sum of entropy across m-body and f-body pixel probabilities is calculated and presented in figure 3.9. The darkest portions of the body of the fly in the image represent the lowest entropy (highest confidence) prediction from the model. Comparing the male and female body-pixel cuts and using a threshold to only keep the pixels with entropy less than 0.5 in figure 3.11 it is clear that the spots and darkened bottom region of the fly are important to the classification decision.

A second example of this is shown in entropy results corresponding to figure 3.12. This demonstration is somewhat interesting because the second wing underneath the fly has been classified as female. If this were isolated it would be difficult to tell whether this wing is male or female because of the lighting conditions and angle that the wing is presented to the camera. This high entropy thin wing is visible in figure 3.13. Making the threshold cut at 0.5 for the entropy, the thin cross-section wing disappears and what is leftover is the wing that is much easier to use for classifying the fruit fly as male. This is shown in figure 3.15. This also may give insight into why the data has a high false negative rate (misclassifying males into a female bin); the wings of the males when they are on the side sometimes do not display the distinct spot and the algorithm may classify those examples as

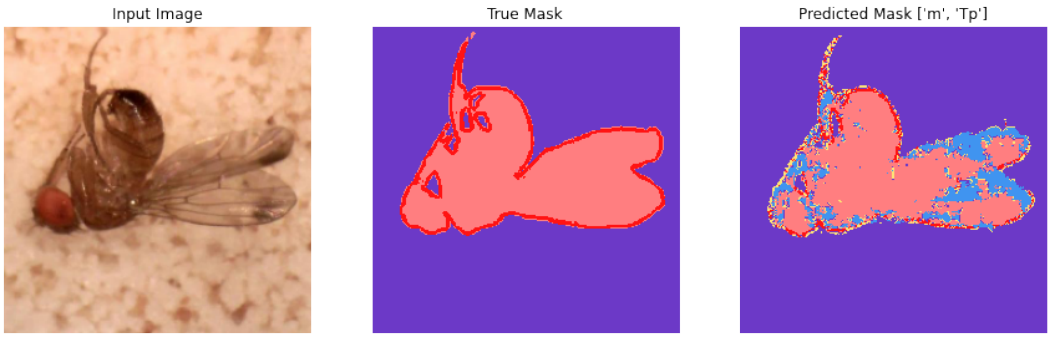

Figure 3.8: (Left panel) A male from the validation set, (middle panel) its human drawn mask, (right panel) the predicted output from the model. For the true masks and predicted masks the light red pixels are identified as male-body pixels, the stark red pixels outlining the fly are male-edge pixels, the light blue pixels are female-body pixels, the yellow pixels are female-edge and the purple pixels are background pixels.

female. In the case of this example the algorithm successfully labels the image as male, since there is another wing available and the darkened bottom is also a key identifier.

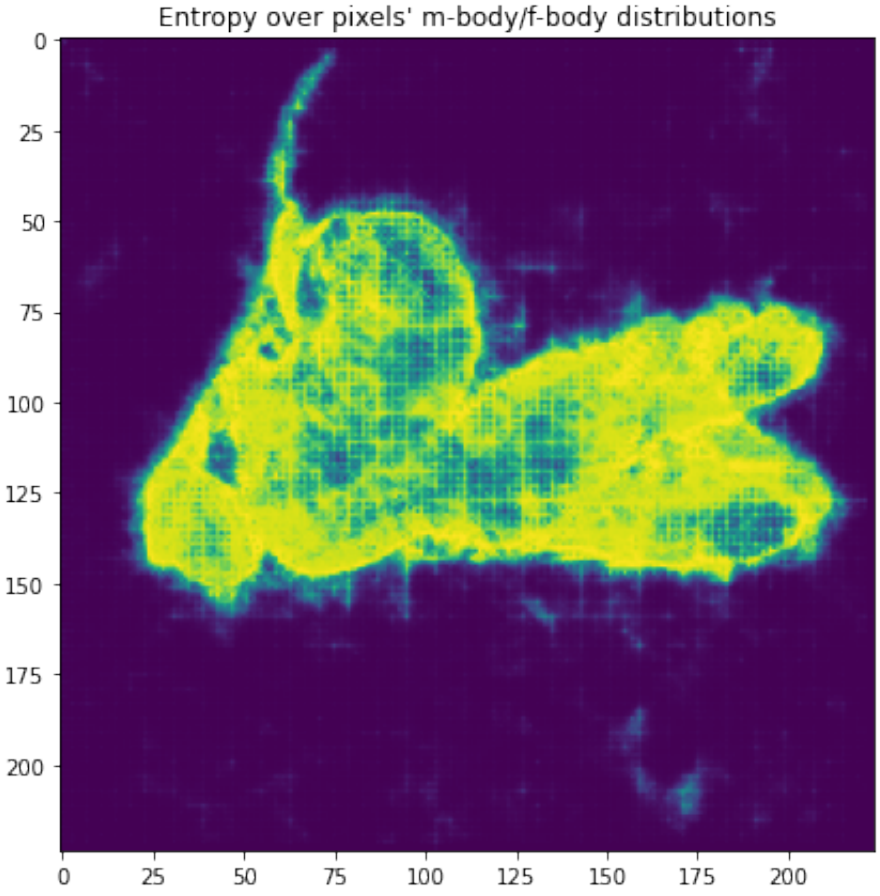

Figure 3.9: A male from the validation set is depicted from its male-female body pixel-wise entropy which is relatively high in most parts of the fly but lowest among the wing-tip and back regions of the fly.

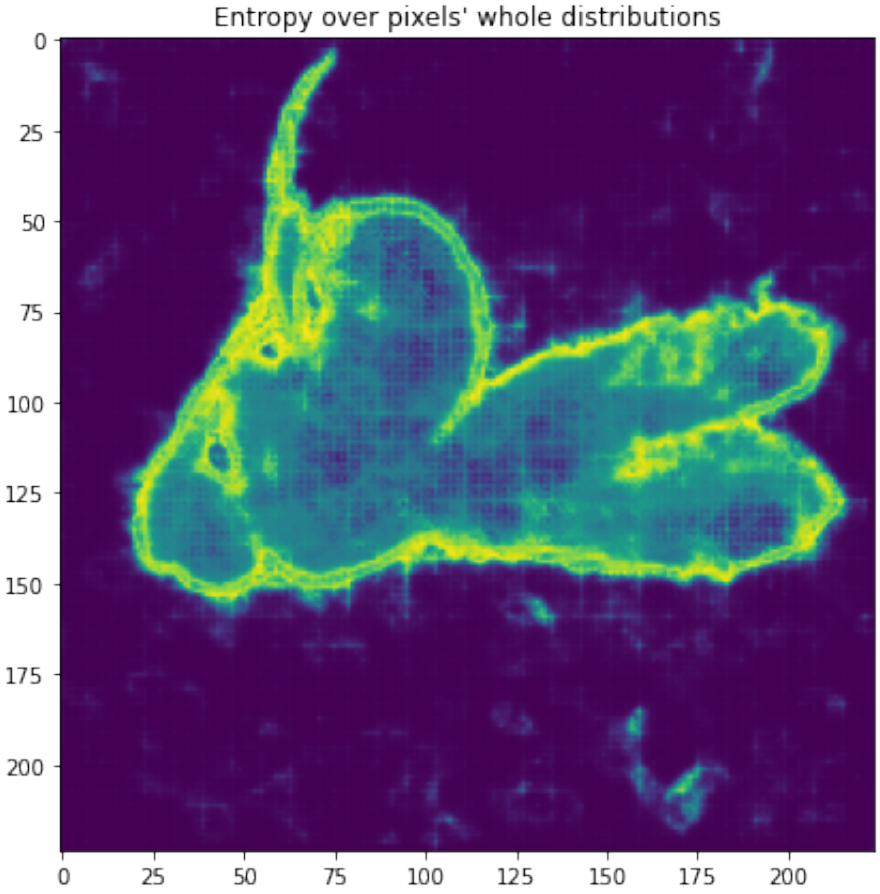

Figure 3.10: A male from the validation set depicted from its entire pixel-wise entropy. The lower entropy regions of the wing-tip and back are still visible, but the contrast is reduced because the edge pixels' entropy are included in the full distribution.

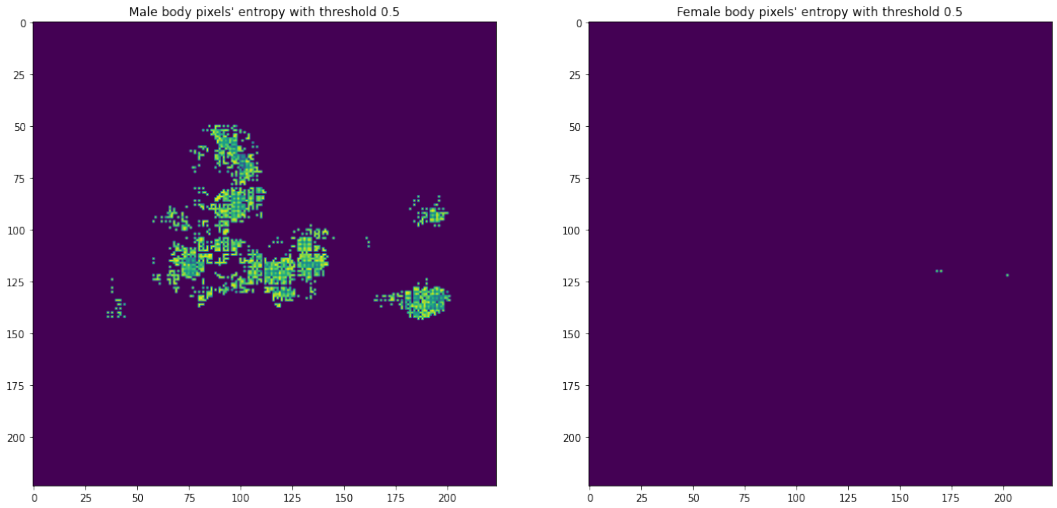

Figure 3.11: A male from the validation set depicted from regions of entropy less than 0.5 for (left panel) male-body pixels and (right panel) female-body pixels. There are almost no pixels on the right hand side, so this is a strong prediction for male. Some of the highlighted regions on the male pixels however are the same regions that are necessary for a human to identify between male and female. The segmentation model has more confidence in morphological features that are more obviously present.

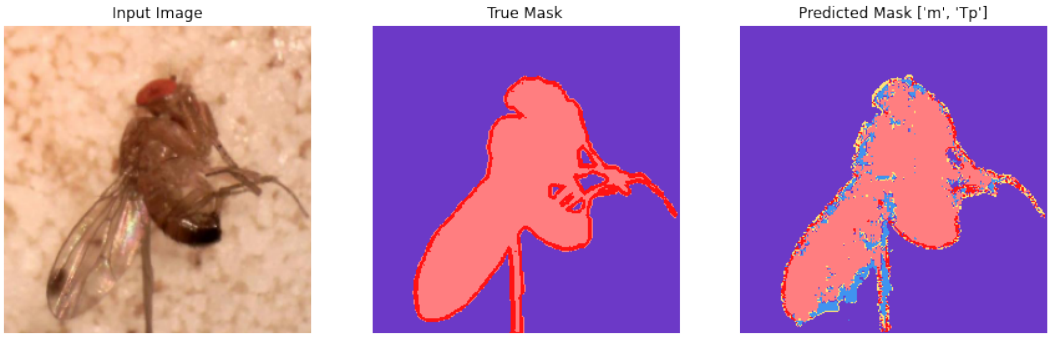

Figure 3.12: A 224x224 male from the validation set which is classified as a true positive. The sideways wing with lower cross-section with respect to the camera angle is mostly classified as female, where the rest of the image is classified as male.

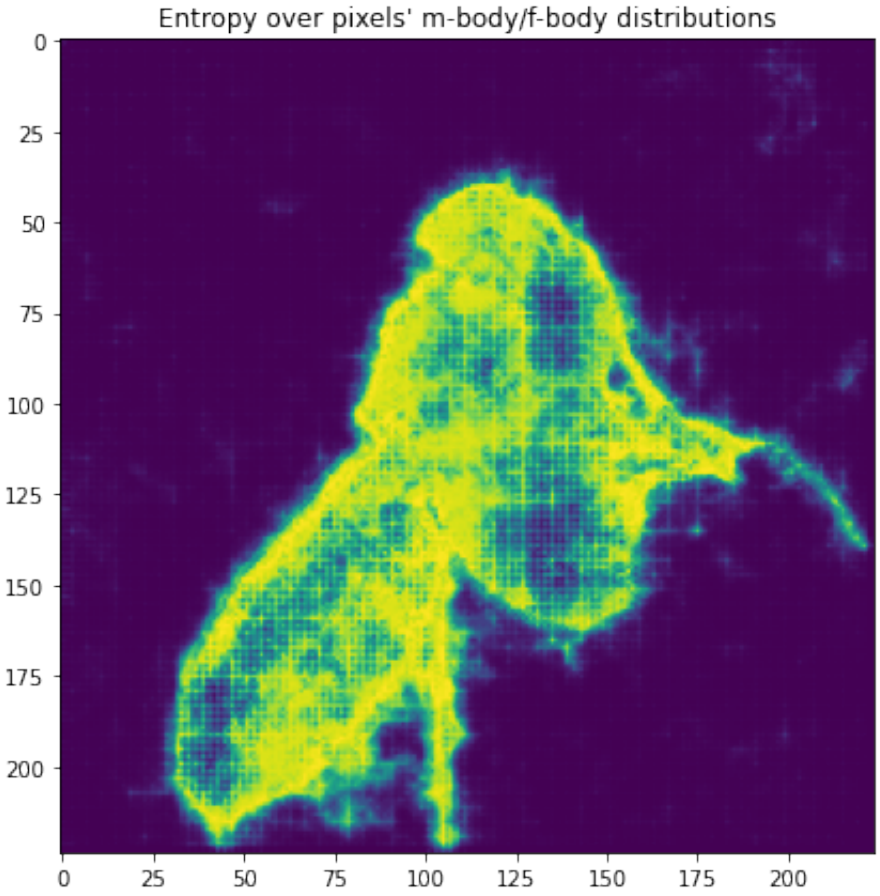

Figure 3.13: The entropy here across male and female body pixels is again lowest along the wing especially by the male spot and also it is low on the bottom and towards the legs.

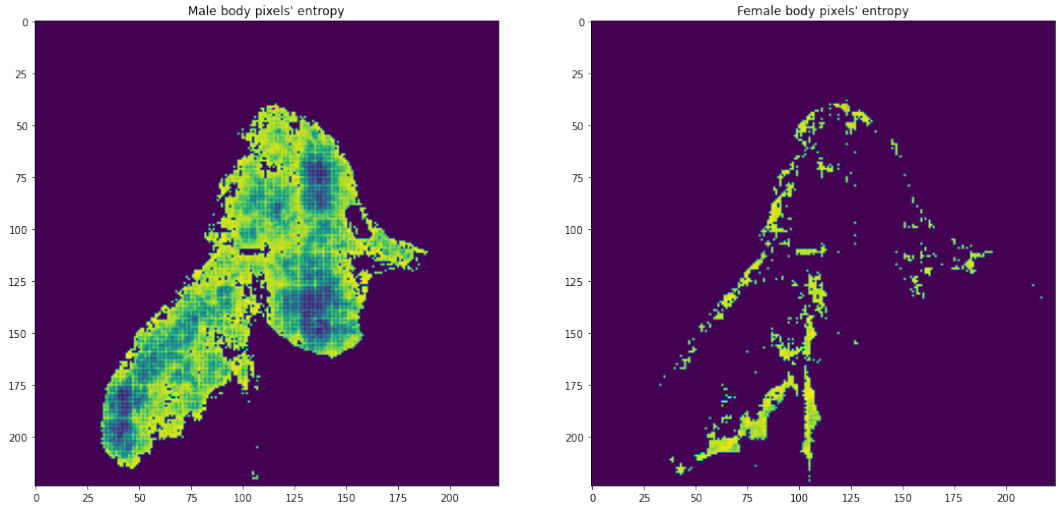

Figure 3.14: The second wing which is thin from this vantage point is completely classified as female, likely because the spots are not as visible with the lighting conditions when the wing is turned on its side.

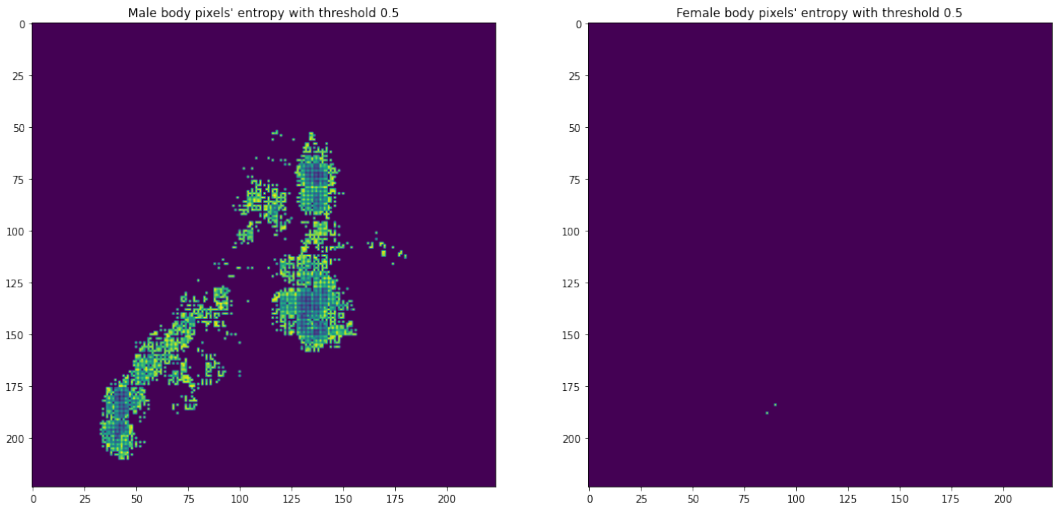

Figure 3.15: Restricting the view to the lowest entropy points below 0.5 shows that the upper wing encapsulating the spot as well as the darkened bottom of the fly are important points for classification.

3.2.5.3 Prediction entropy for counts of body and edge pixels

In the previous section the entropy localized on each pixel was calculated to show the confidence of a model's predictions and how it depends on the morphological features of SWD which are visible in the input image. In this section Entropy for binary predictions based on pixel counts are described, where each set of prediction pixels from an input image are turned into one entropy which quantifies an aspect of the model's overall prediction between male and female. The objective is to see how the confidence of the model operates on the scale of the whole validation set.

To form a probability that acts as a proxy input into entropy, the output from the model must be interpreted so that a decision can be made whether the input image is male or female. While choosing the lowest pixel-wise entropy predictions from body pixels is sometimes a more accurate strategy in classifying between male and female, the objective is to see how much of the image may be confused between male and female. Therefore a simple collection of the male classified edge and body

pixels compared to the female classified edge and body pixels will give a rough idea of the overall output picture. Letting the proxy probability p_i be the number of pixels n_i belonging to class i contained in the image divided by the total number of pixels $N = \sum_i n_i$ from the classes considered in the image so that $p_i = n_i/N$, the entropy is calculated between the male and female predictions:

$$E = \sum_i p_i \log p_i.$$

Since only male and female edge and body pixels are considered (and not background pixels) there are 4 categories instead of 5 for the sum. The number of background pixels varies as a function of the orientation of the fly (assuming a well-trained model) and will make male and female comparisons more difficult if it is included in the estimation of the model's confidence. That is not to say it wouldn't be useful for other purposes, such as taking size of the fly and orientation into account, or estimating other qualities about the prediction that may occur for example during out-of-distribution inputs unseen by the classifier. For the purposes of this analysis only 4 categories are considered, so the maximum entropy that is possible to observe is $\log(4) \approx 1.386$, though such balance is rarely encountered.

For multiple experiments this entropy is calculated and displayed under evaluation from multiple different sources of data, including the validation data set, a small set of challenging samples, and the leftover data of medium difficulty combined with the OOD data. A histogram of the entries in the confusion matrix is produced which places instances of true-positive (Tp), true-negative (Tn), false-positive (Fp) and false-negative (Fn) results into bins of entropy.

3.2.6 32x32 experiments on SWD

This is an experiment with size 32x32 images of flies which examines the affects of augmentation on the SWD data sets. Using a smaller resolution degrades some of the detailed features and makes color-based features more important. The deleterious effects from augmentation which are unintended may be less relevant at this scale so that the confidence of the model's prediction can be examined under the affects of augmentation without degrading the accuracy. Another advantage of the 32x32 size images is that more of them are able to be held in memory on a Google Colab instance for detailed analysis. Improvements utilizing persistent storage or making analysis without storing as much in RAM could improve this minor implementation issue.

For this experiment 32x32 size images were used in an experiment to train two models out to the same number of epochs (180 epochs). One of the models did not augment its training data and the other augmented virtually every other sample. There was a 10% probability of applying each augmentation sequentially to each input image. Approximately 47% of the data encountered augmentation during training of the augmented model.

3.2.6.1 Results of 32x32 experiments without augmentation

These are the results of the 32x32 experiments without augmentation. The segmentation model performed relatively well given the small input size of the data. Using the heuristic binary evaluation of number of male to female pixels as the decision criterion, the model maxed out at about 85% accuracy at 75 epochs before decaying to about 83% accuracy at the final epoch, which where the predictions for

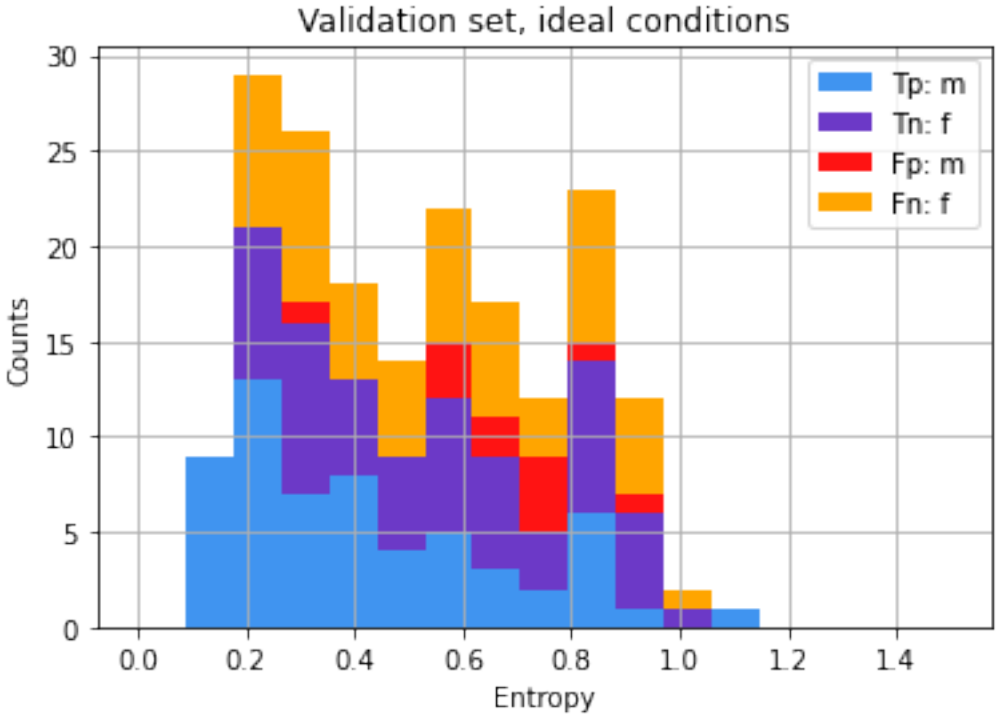

Figure 3.16: 32x32 confusion matrix and entropy histogram from a model trained on unaugmented data. This is from the validation set, which contains data that is most like the training data and easiest to use. There are as mentioned previously quite a few false negatives. Notice how the false negatives are relatively uniform across the entropy, but the false positives are more heavily centered towards high entropy. This shows that the validation data chosen for this run has mostly low entropy samples, but those samples which are false positive are also high entropy. This means that when the model is confident about male predictions, it's more likely to be right. For negative predictions this does not hold, confident negative predictions still have a high false negative rate (meaning males are lost, presumed to be females).

analysis come from. Not bad considering the features are barely visible with 32x32 images. This shows that coarse color and coarse contrast are useful classification properties for SWD.

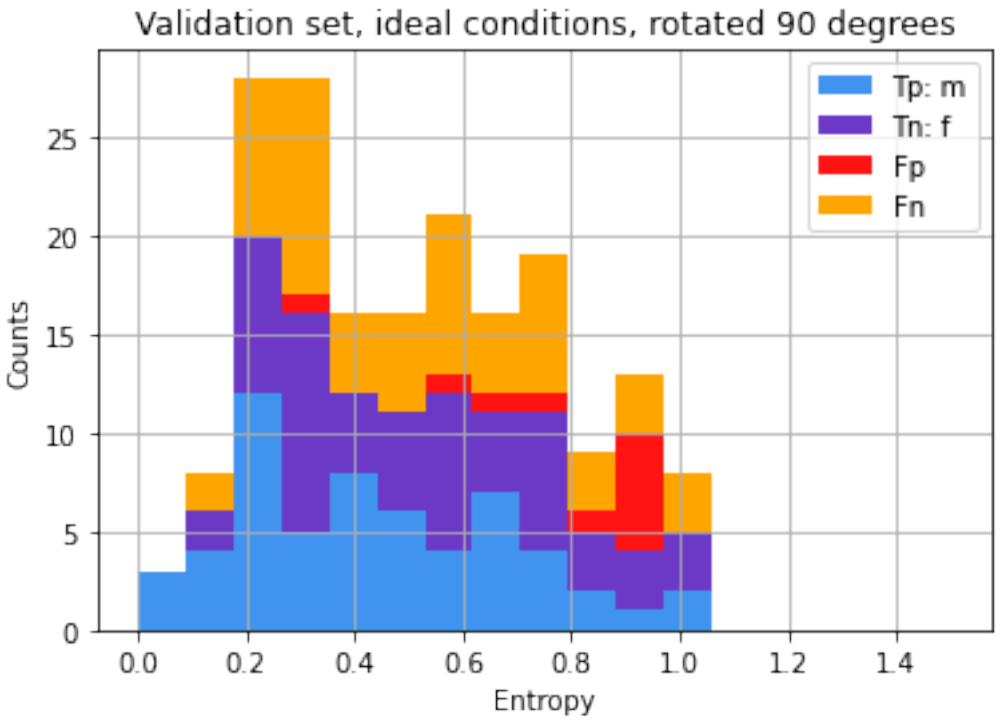

Figure 3.17: 32x32 confusion matrix histogram from a model trained on non-augmented data. To re-evaluate this validation set each image is rotated by 90 degrees and the histogram is calculated again. This gives similar results to the previous histogram with values changed not so drastically. Notice how the false positives are mostly distributed along high entropy still. Most of this validation data has low entropy again. The peak is roughly between 0.2 and 0.3.

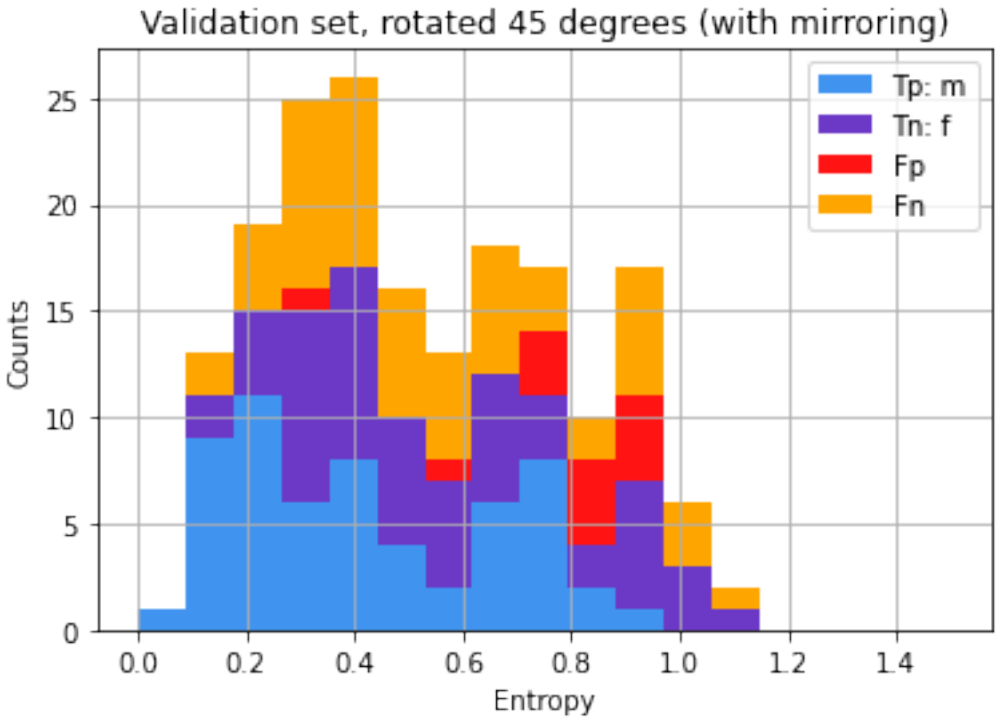

Figure 3.18: 32x32 confusion matrix histogram from a model trained on non-augmented data. One of the types of rotation that should introduce additional entropy is rotation with "mirroring", which rotates a square image and fills the newly vacant corners of the image with reflected values from the rotated image. This is a common way to create augmented data for training. Notice how the peak of the entropy has shifted to the right all the way up to 0.4. The entropy of the whole data set has slightly increased when rotated at an angle where mirroring is a large factor.

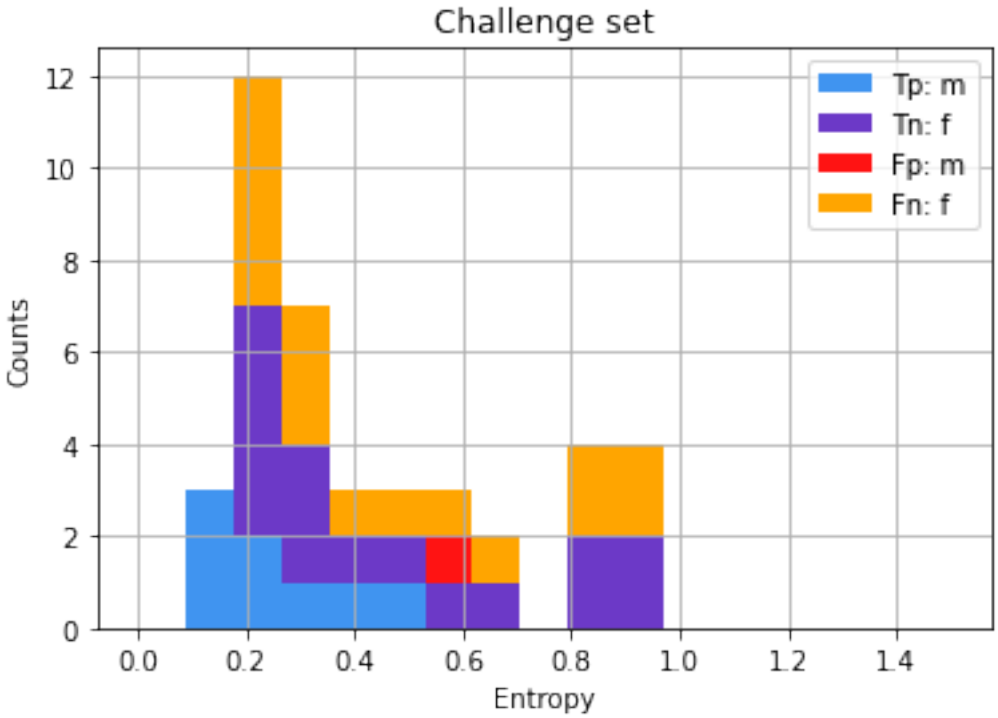

Figure 3.19: 32x32 confusion matrix histogram from a model trained on non-augmented data. This data was from a small set of examples called the "challenge" or "difficult" set in figures. This was because it was difficult for the labeler to identify whether the sample was positive or negative. Consistent with the results of training, the false negative rate is high and the entropy of these samples is relatively low.

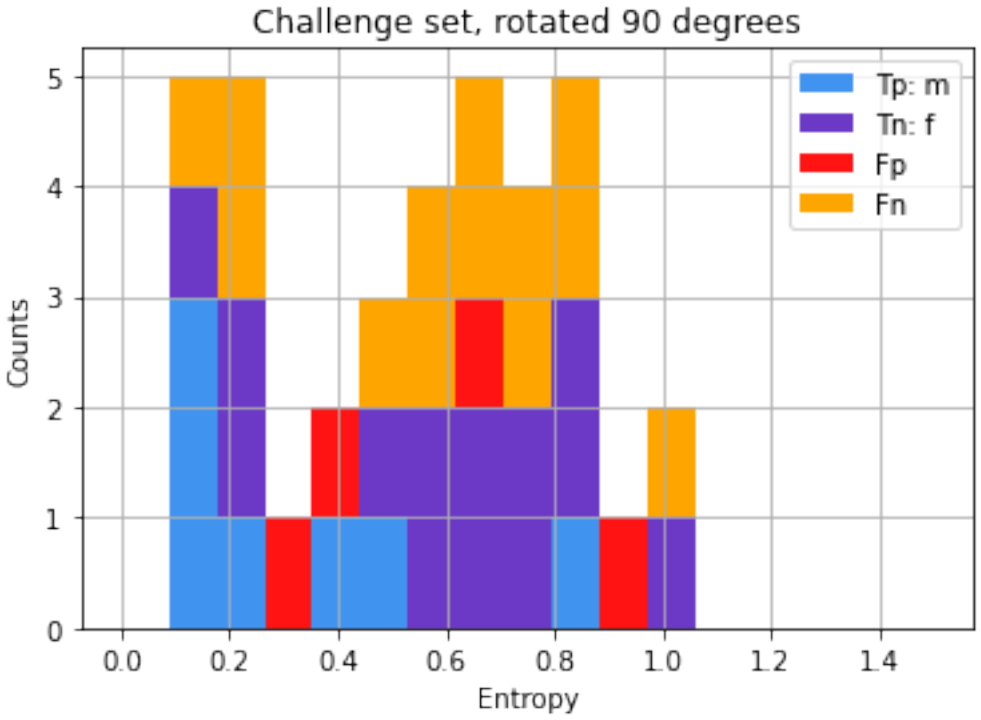

Figure 3.20: 32x32 confusion matrix histogram from a model trained on non-augmented data. When the challenge set is rotated by 90 degrees many of the false negative results move up to a higher entropy and now there are a few more false positives. This variability from a small magnitude rotation (no mirroring for 90 degree rotations) may indicate that these samples are difficult, though the sample size is quite small.

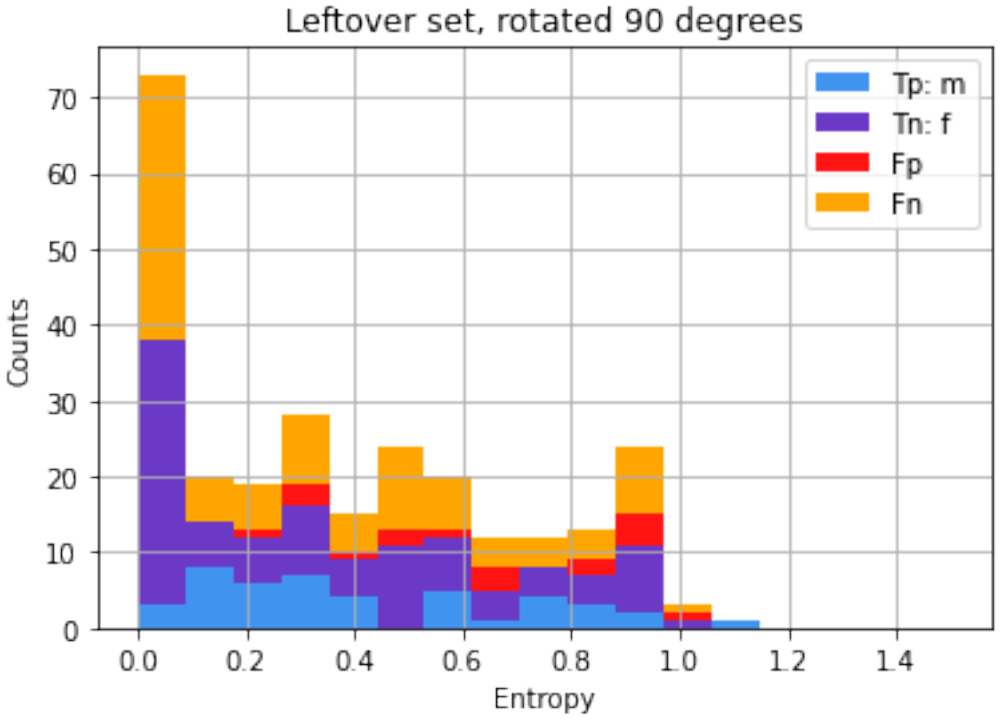

Figure 3.21: 32x32 confusion matrix histogram from a model trained on non-augmented data. This is the leftover set which includes examples of medium difficult which are similar to the training data set as well as the out-of-distribution examples where contain many different colors not seen in the training data set. Since the sample size is larger it is easier to see quantitative changes in the entropy due to mirroring. The spike of super low entropy false negative and true negative samples are indeed the OOD set. The model completely whiffs on the OOD set at 32x32 resolution and classifies them all as female. The rest of the well-behaved data is visible with a relatively uniform split between false and true examples across the entropy, though the false positives still have higher entropy.

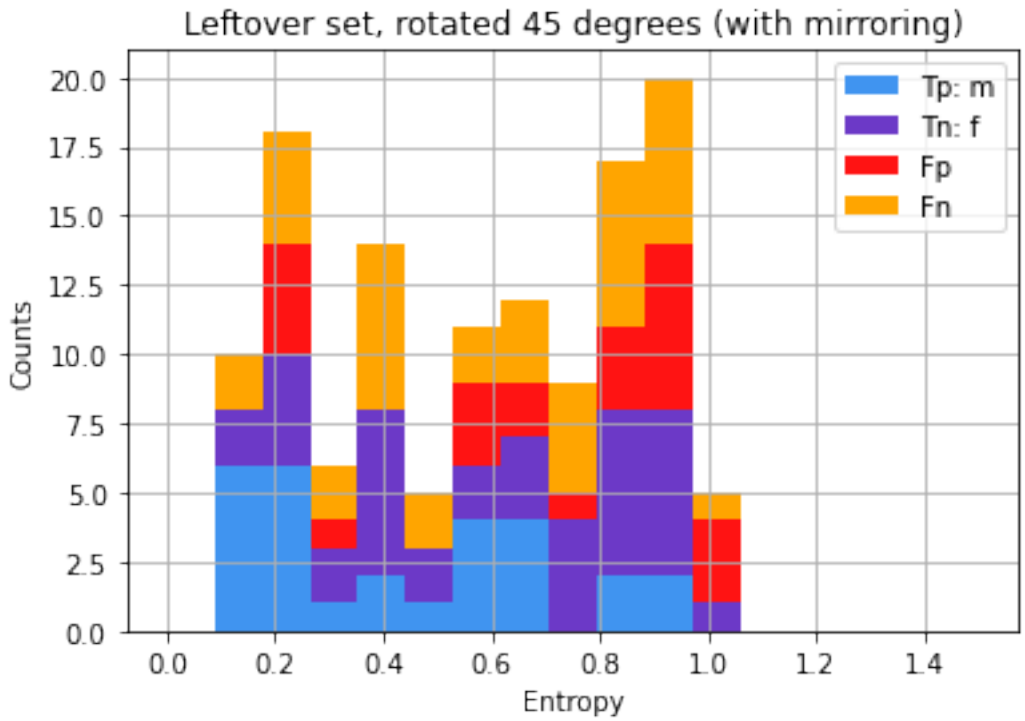

Figure 3.22: 32x32 confusion matrix histogram from a model trained on non-augmented data. When rotation with mirroring is tested on this same leftover set, it is clear that the entropy increases quite a bit across the data set. The change in distribution of the data is even strong enough to break the high entropy false negative bias exhibited on the OOD data from the previous figure.

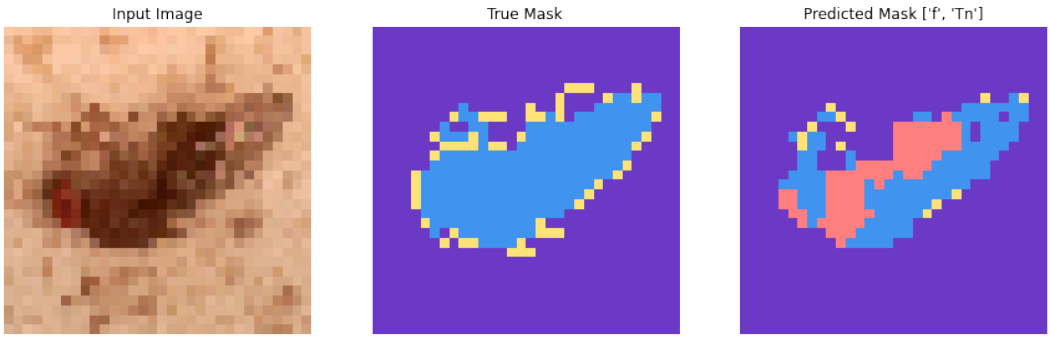

Figure 3.23: 32x32 prediction example for a female SWD. The fly is split in prediction between male and female but the wings play a major role in the correct classification. This is one of the examples of the low resolution experiments where the wing was clear enough that the algorithm was able to make an identification.

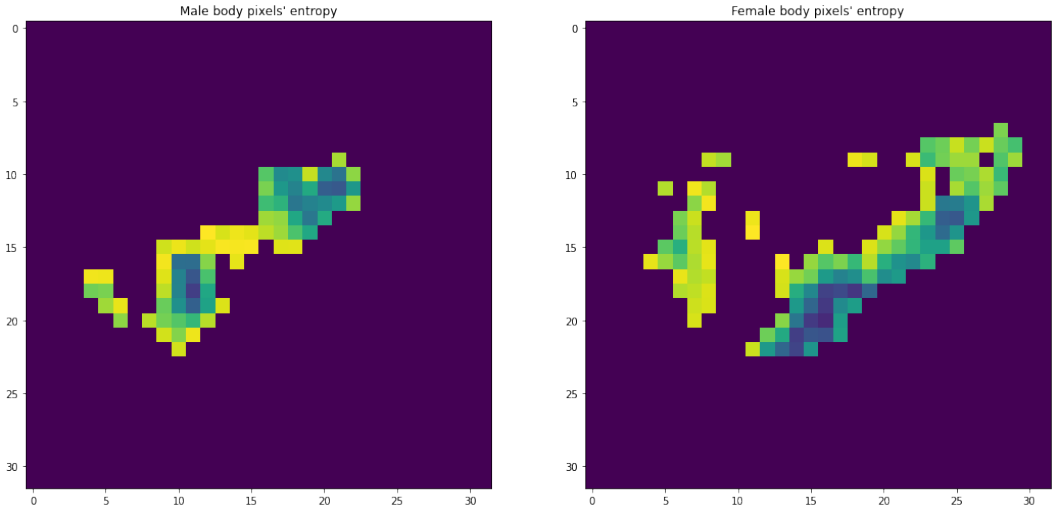

Figure 3.24: 32x32 female image showing the lowest entropy along the wing of the female.

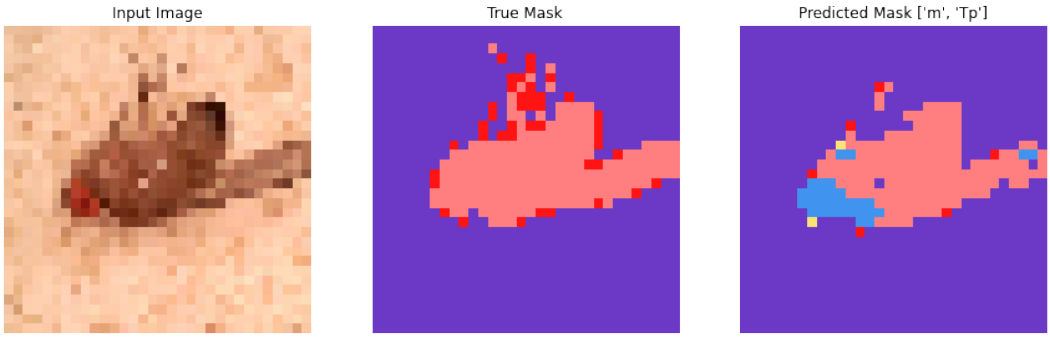

Figure 3.25: 32x32 prediction example for a male SWD classified as a true positive. The male is on its side and the spot on the wing is slightly less visible. With the low resolution it may be difficult to tell whether this is male or female with the untrained eye. The bottom of the male allows for positive identification as it is quite dark.

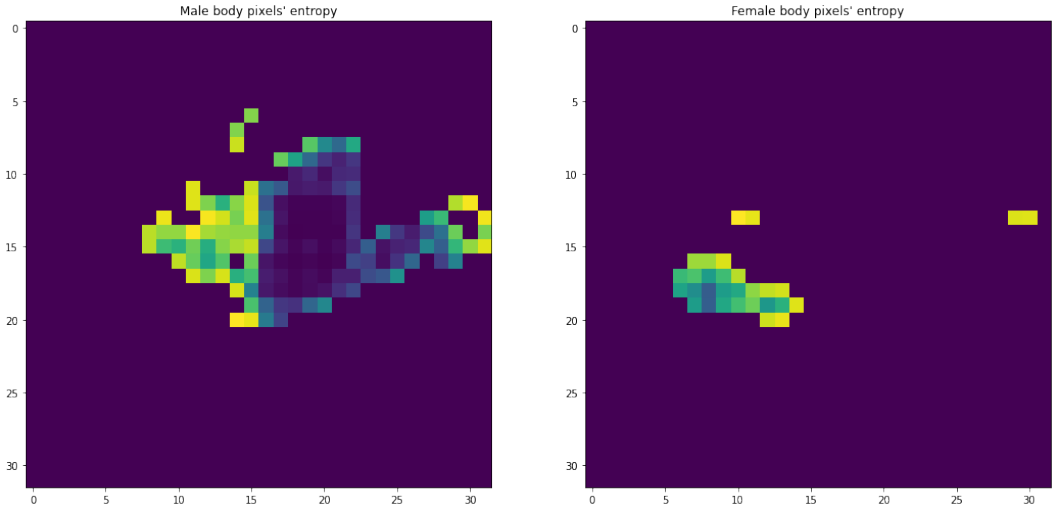

Figure 3.26: 32x32 male shows the lowest entropy for the male is along its back half with the dark bottom. The low entropy extends out into the wings but becomes higher as the wing gets closer to the edge of the image. Images where the tip of the wing get cut off may be responsible for this increase in pixel-wise entropy.

3.2.6.2 Results of 32x32 experiments with augmentation

These are the results of the 32x32 experiments with augmentation applied to the model during training. This model also performed well with the heuristic binary evaluation during segmentation training. The binary accuracy peaked above 87% 137 epochs in and decayed to 81% out to the final epoch used at test time. It should be expected that the augmentation increased the number of epochs it takes to reach the peak in accuracy, since there are effectively more training examples to use and more variability to train on. The same histograms are viewed with this augmented model trained out to 180 epochs.

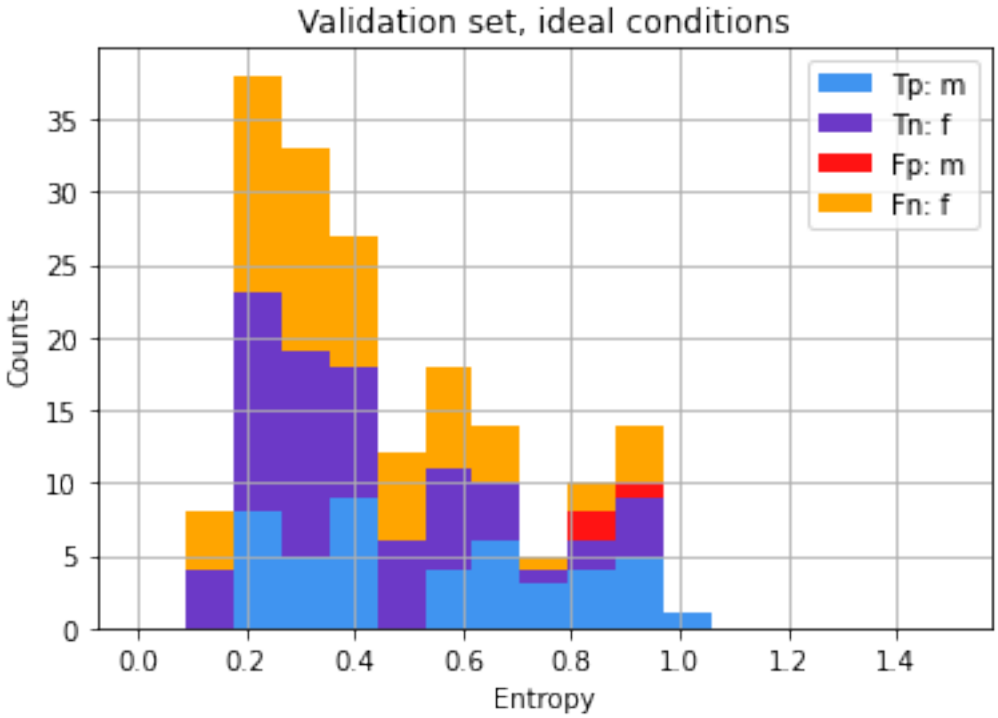

Figure 3.27: 32x32 confusion matrix histogram from a model trained on augmented data. Again the training results in a relatively low false positive rae and high false negative rate. The main observation in comparing this figure to **??** is that the average entropy has gone down for most of the samples. This will be examined more in the following figures.

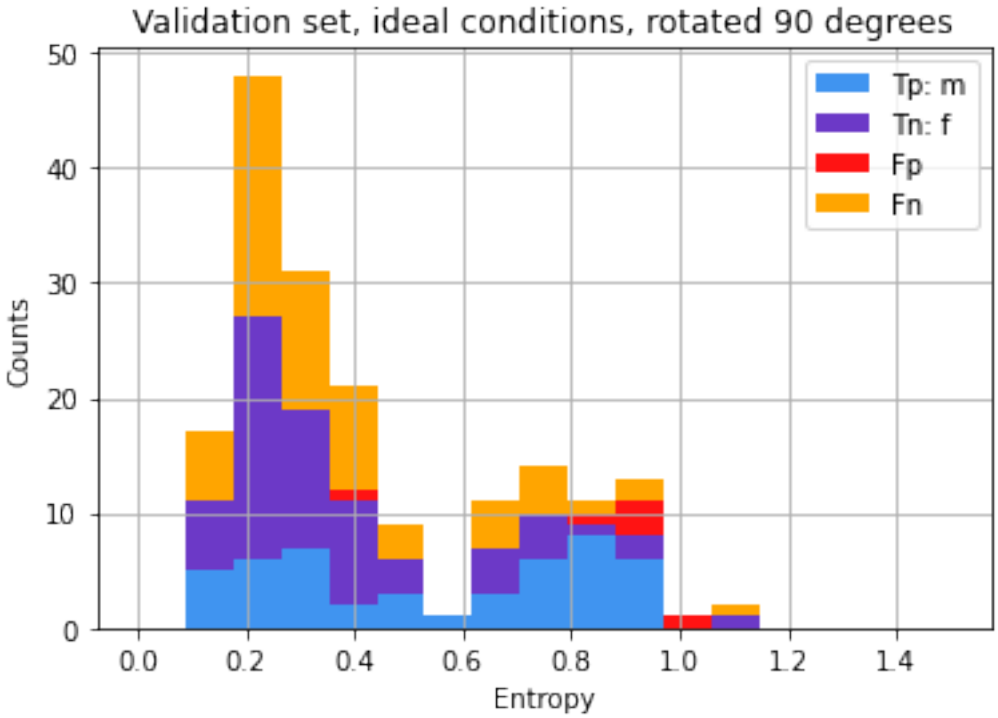

Figure 3.28: 32x32 confusion matrix histogram from a model trained on augmented data. Rotating the validation set by 90 degrees has appears to have very little effect on the distribution of entropy, perhaps less so than the non-augmented 90 degree rotation had. In comparison to the non-augmented model's validation data this appears to have quite low entropy.

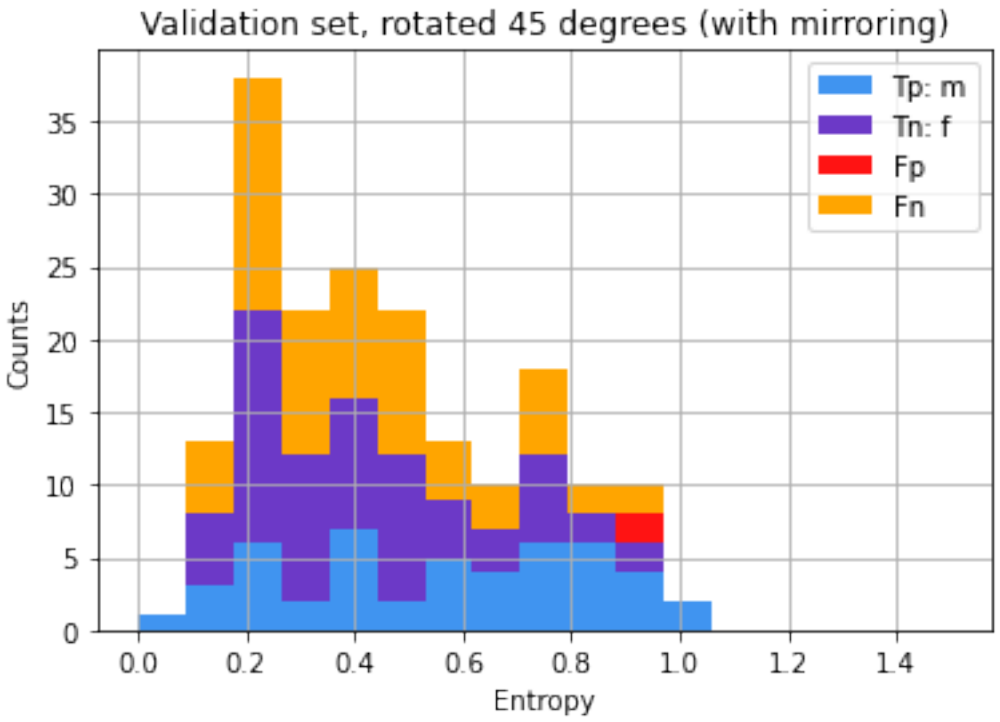

Figure 3.29: 32x32 confusion matrix histogram from a model trained on augmented data. Rotating the validation set 45 degrees so that mirroring has an effect on the data, an increase in the entropy of samples is seen again, but perhaps smaller in magnitude. The peak has decreased in height but not in location as it did with the non-augmented model. This could be construed as evidence that the augmentation successfully reduced the entropy increasing effects of rotation with mirroring.

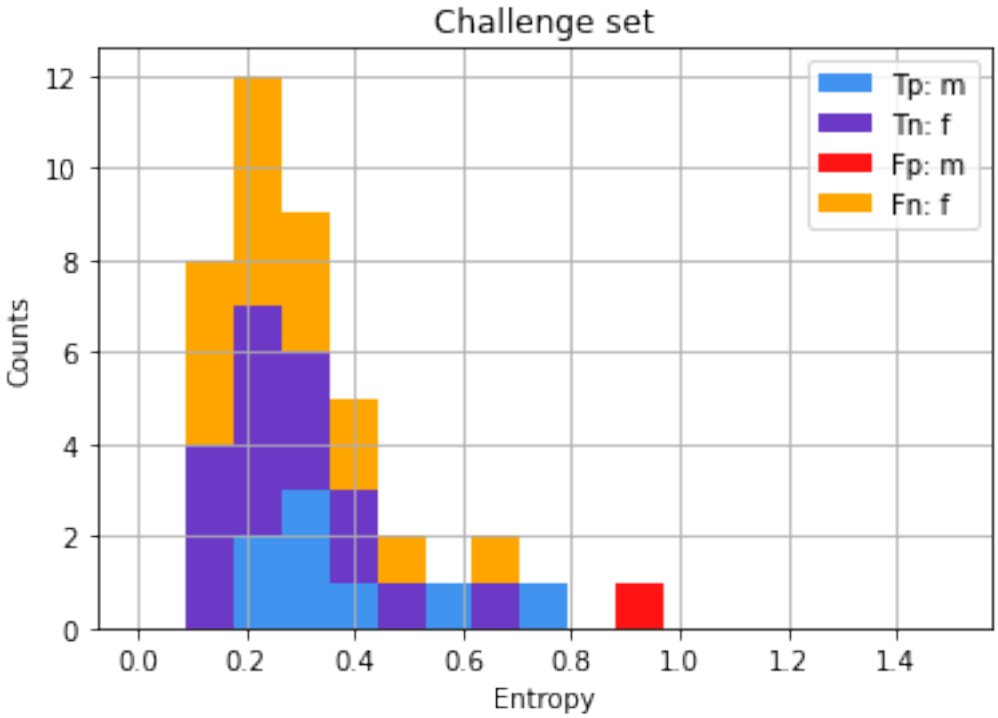

Figure 3.30: 32x32 confusion matrix histogram from a model trained on augmented data. The distribution of entropy on the examples from the challenge set show a very similar distribution, though again the entropy has decreased. Strangely, rotation of the challenge set by 90 degrees increases the entropy again, though not by as much as it did for the non-augmented model.

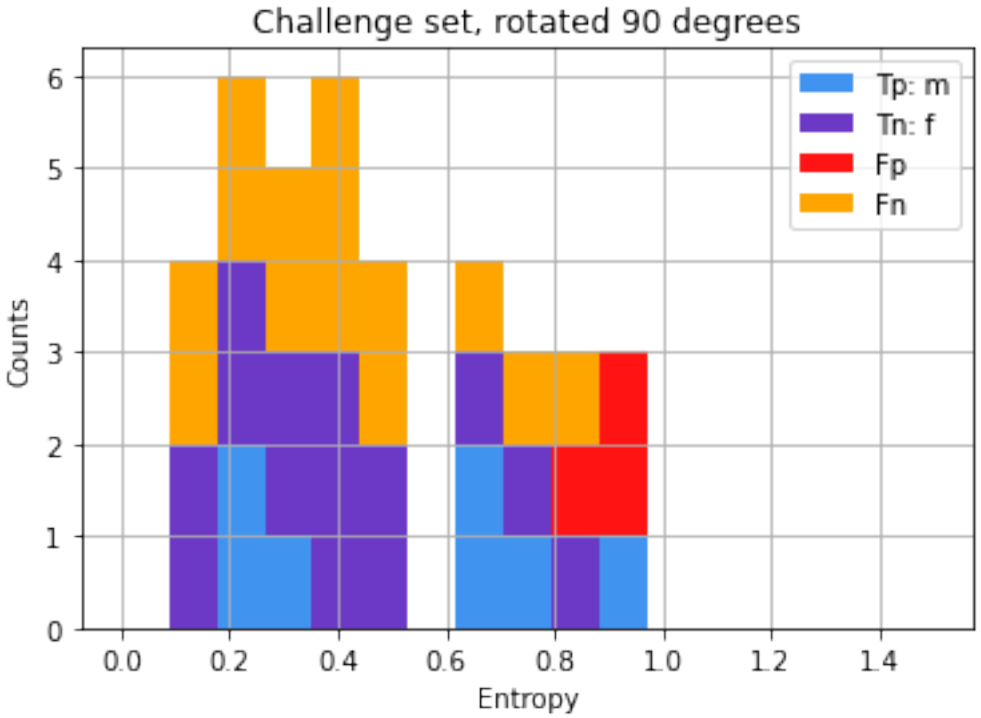

Figure 3.31: 32x32 confusion matrix histogram from a model trained on augmented data. Viewing the leftover set the spurious female bias of the model on data from the OOD set is exhibited again alongside the regular data. The shape of the entropy a sloshed towards the left, indicating a lower average entropy again for the augmented model in comparison to the same results for the non-augmented model.

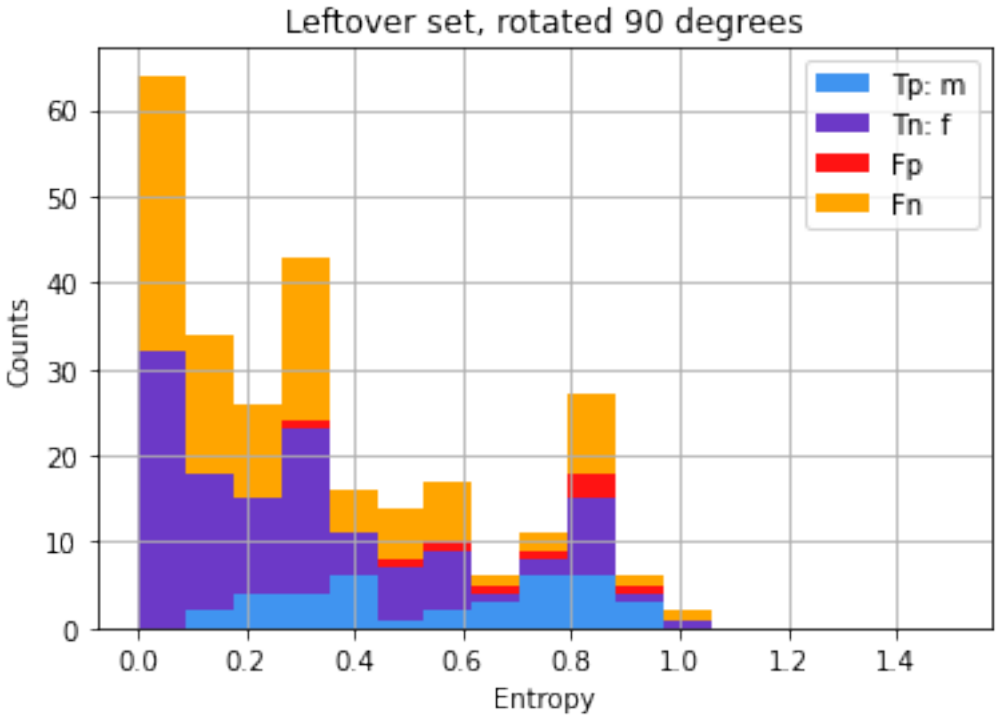

Figure 3.32: 32x32 confusion matrix histogram from a model trained on augmented data. In comparison to the non-augmented data again the entropy for most examples has moved to be smaller than before, though it appears that some of the missclassified OOD examples moved into higher entropy bins (there are 76 OOD examples which were mostly in the first bin of entropy in figure 3.21).

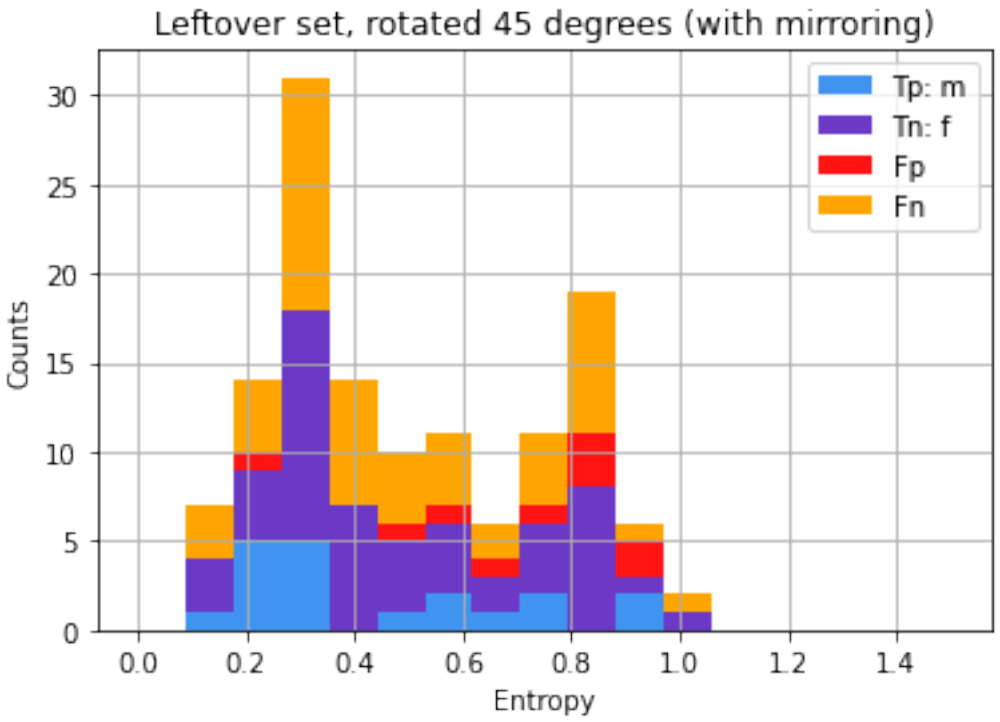

Figure 3.33: 32x32 confusion matrix histogram from a model trained on augmented data. When the leftover set is rotated at 45 degrees so that the mirroring effect is present again, the entropy increases and the false positive rate increases too, but not nearly as much as it does for the non-augmented model.

3.2.7 128x128 experiment on SWD

This is an experiment with size 128x128 images of flies which examines a few cases of the morphological characteristics which can lead to positive and negative identification based on the entropy of individual pixel predictions. Examination of pixel-wise entropy is presented for in-distribution validation examples of 17x and 27x magnification, as well as OOD examples with 10x magnification. The heuristic binary accuracy during training of this model ended just below 80% as it was stopped at 120 epochs, which is earlier than even the small 32x32 models. Nonetheless it produced results consistent in identifying morphological features like the models from the previous sections.

3.2.7.1 Observations of important regions on SWD on laboratory examples

The following examples demonstrate the model's predictions on a male fly from the 17x magnification set and a female fly from the 27x magnification set. Both examples are correctly classified; however the low entropy regions on the male more distinctly point out a successful morphological identification than they do on the female. On this particular experiment's run the model's training was ended on an accuracy decrease where specificity slightly decreased and recall slightly increased. Perhaps a portion of the low entropy pixels classified for the female in this section were a result of temporary over-fitting. Still other true negatives not presented here had regions of low entropy which did capture regions near or just behind the ovipositor of the female. For many examples still that region of prediction on the fly along with the spotted wing region is reserved for features more sensitive to high

Figure 3.34: 128x128 prediction for sideways male at ≈17x magnification. The model was able to identify this example, and despite the fact that the fly is sideways the spots on the wings are visible this time.

confidence in male detection, even in some females.

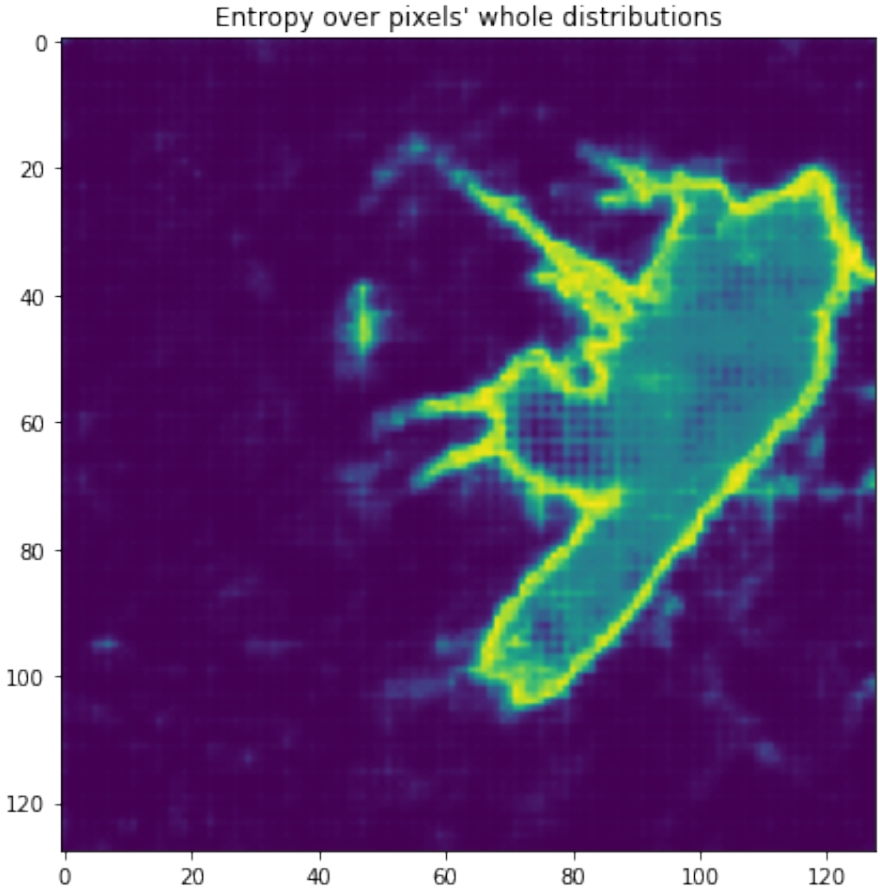

Figure 3.35: 128x128 pixel-wise entropy over the whole distribution for the male SWD at ≈ 17x magnification. Since the background and edge predictions for each pixel are included in the whole distribution, the edges are visible. The darkest parts on the fly are the lowest entropy points usable for classification of male and female.

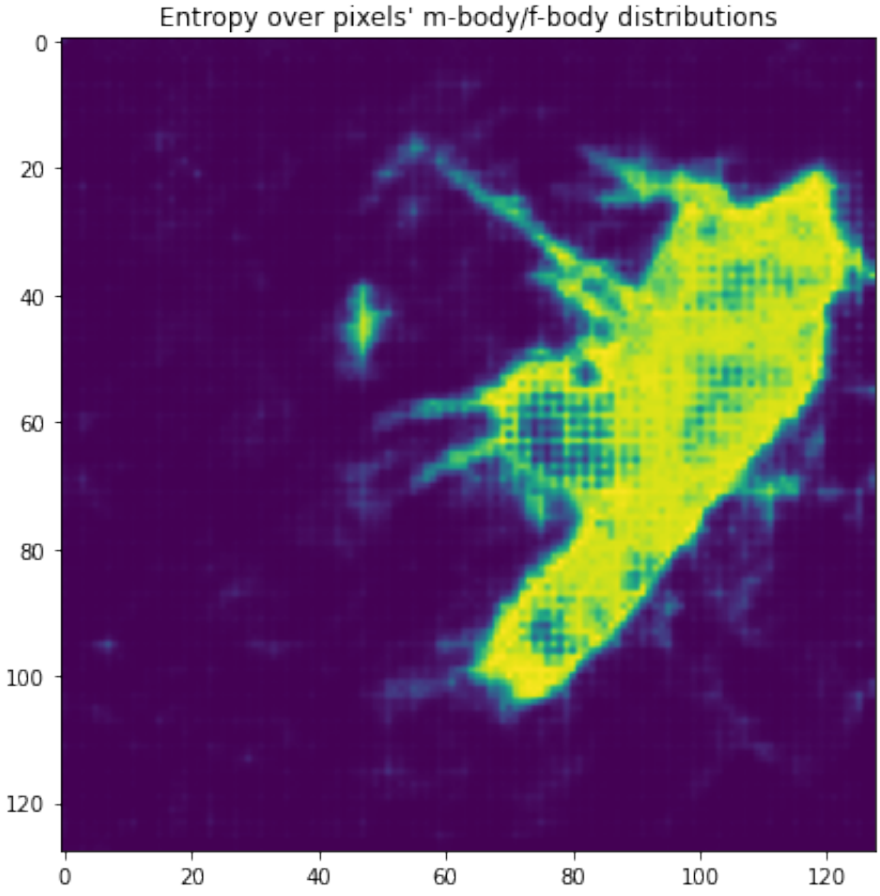

Figure 3.36: 128x128 pixel-wise entropy over only the male and female body pixels for the male SWD at $\approx$ 17x magnification. The contrast between the bright and dark places is now less impeded by the high entropy edge pixels.

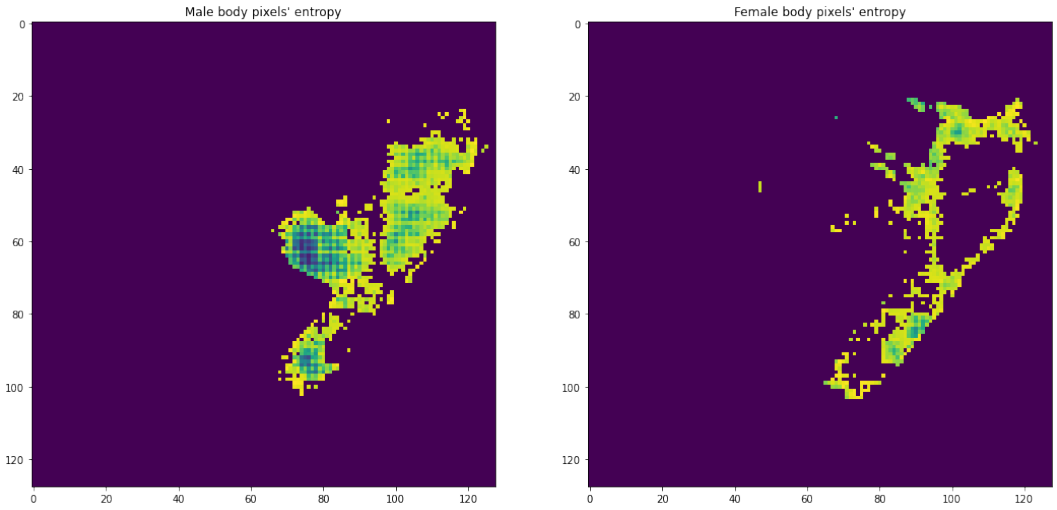

Figure 3.37: 128x128 pixel-wise entropy over the whole distribution except cut into regions of male body identified or female body identified pixels for the male SWD at $\approx$ 17x magnification. The spot on the wing and the darkened bottom on the male are clearly the most important parts responsible for male classification.

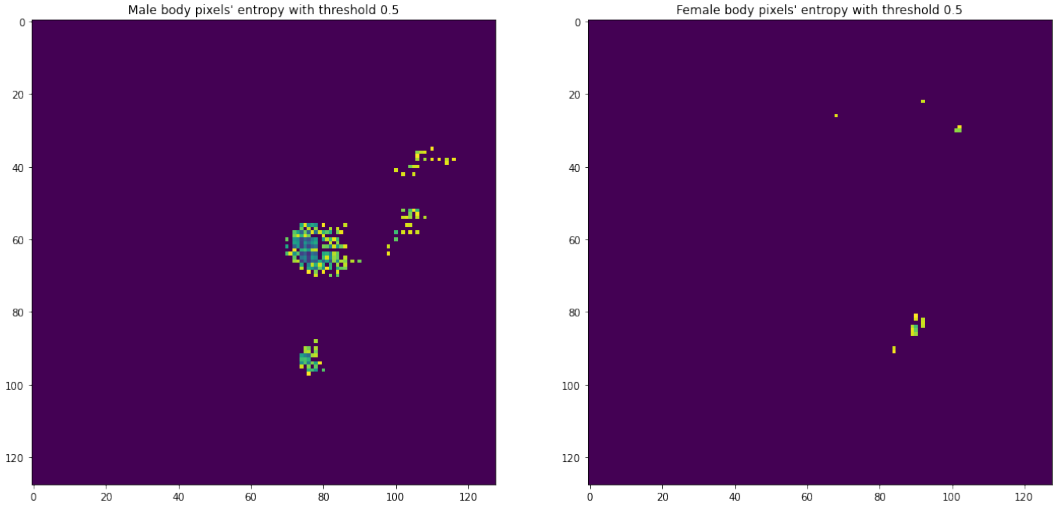

Figure 3.38: 128x128 pixel-wise entropy over the whole distribution except only for entropy below 0.5 and cut into regions of male body identified or female body identified pixels for the male SWD at $\approx$ 17x magnification. The spot on the wing and the darkened bottom are more specifically highlighted.

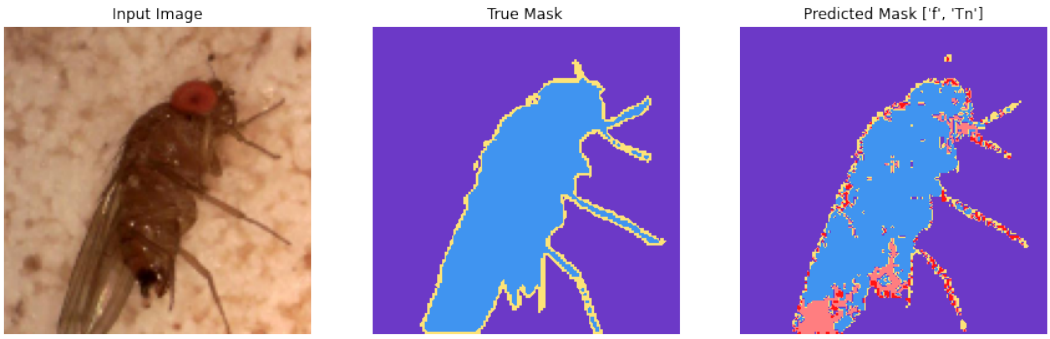

Figure 3.39: 128x128 prediction for sideways female at ≈ 27x magnification. The serrated ovipositor is clearly visible and there is no spot on the wings; however for this example the tip of the wing is cut off.

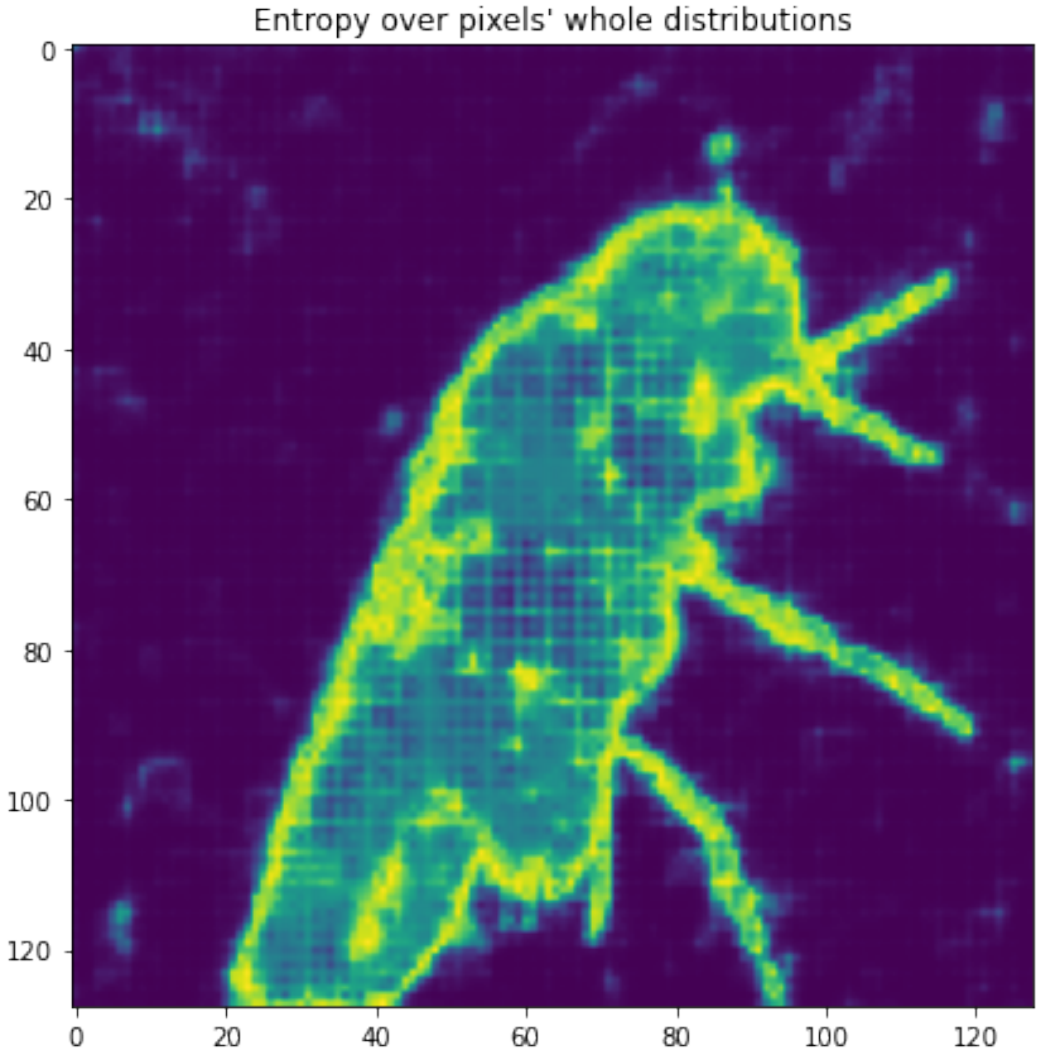

Figure 3.40: 128x128 entropy for sideways female. There is no clear region which is immediately obious as the darkest point. The bottom of the fly and tips of the wings do not have very low entropy.

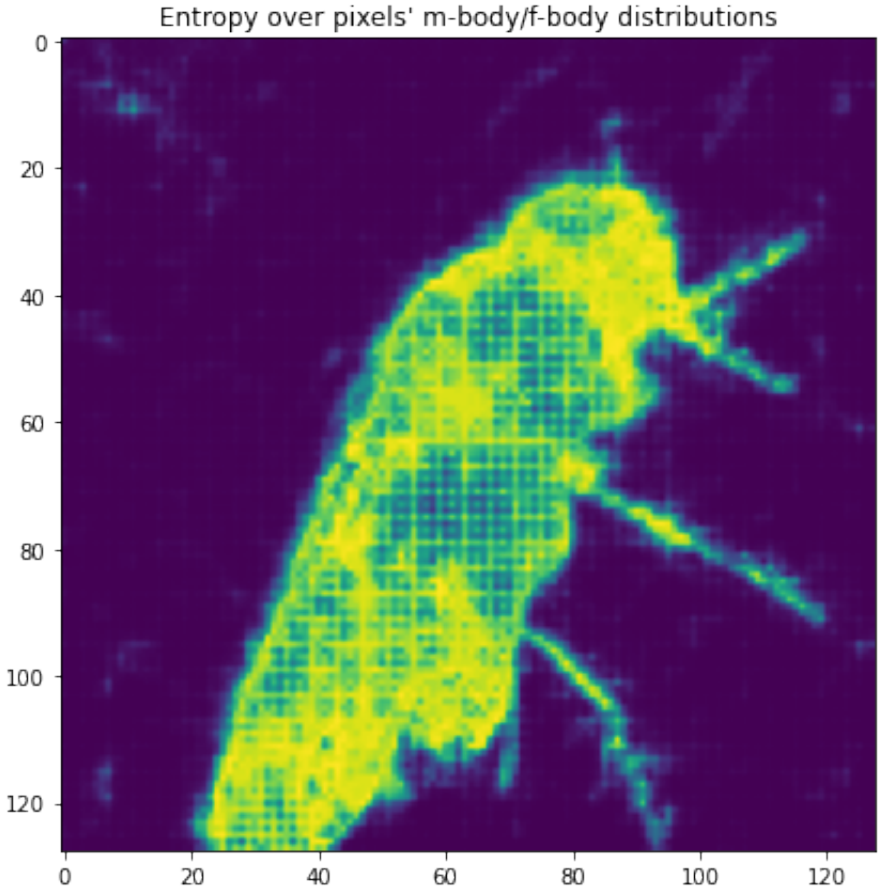

Figure 3.41: 128x128 prediction for sideways female. It becomes clearer that the lowest entropy points are towards the center of the female. The bottom and wingtips are not strong indicators for the model, but rather the central region of the fly, which may be an indicator of some tendency of this model to erroneously hone in on the center of the fly (possibly due to he rotation augmentation not utilizing translation).

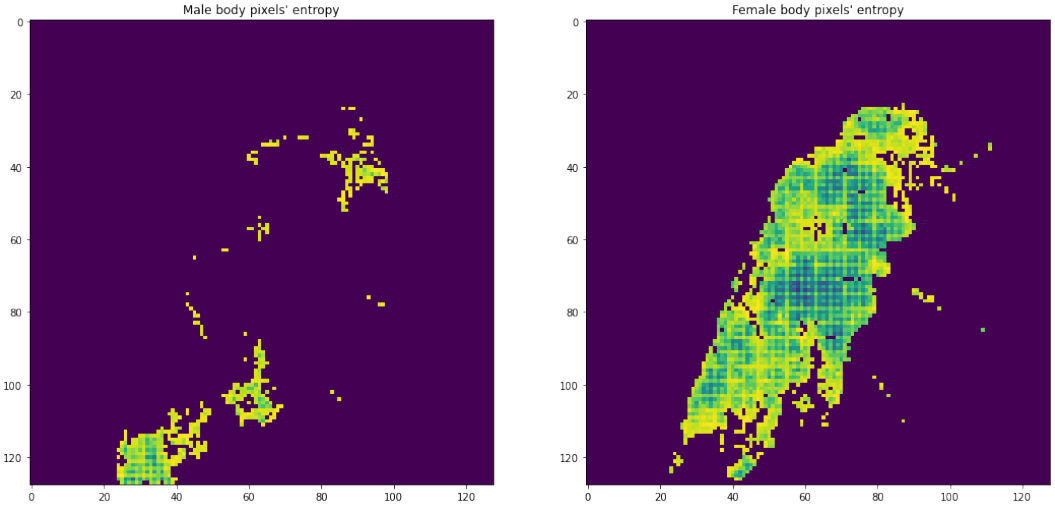

Figure 3.42: 128x128 prediction for sideways female. The pixels which have entropy lower than 0.5 are mostly on the female side of the prediction. Wing pixels toward the edge of the image are still misclassified as male pixels.

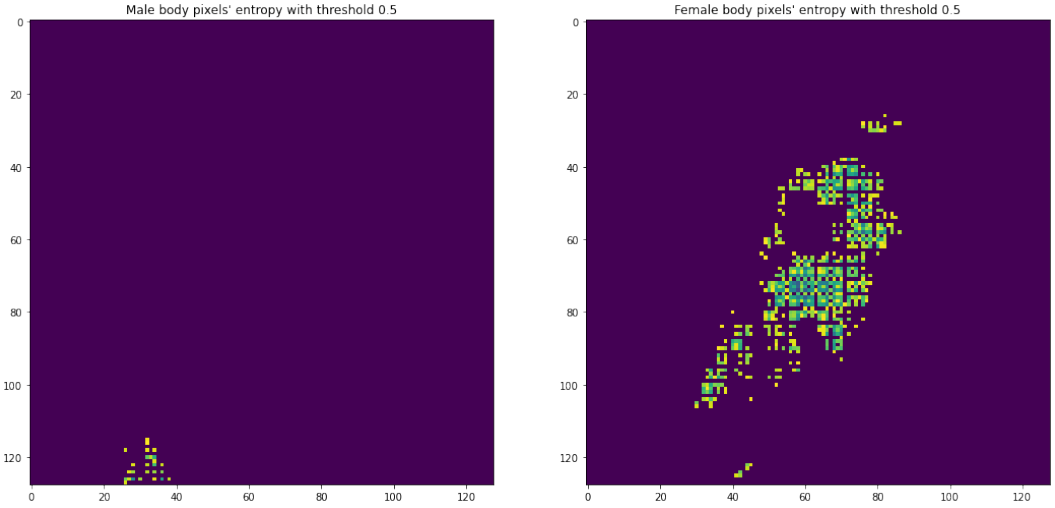

Figure 3.43: 128x128 prediction for sideways female. While the center-top of the fly is an important part of the classification, pixels along the darkened transparent top parts of the wing and a few near the serrated ovipositor make the cut.

3.2.7.2 Observations of important regions on SWD on OOD examples

In this section presentation of a true positive and true negative captured from the OOD set will show that the model does retain some ability when trained on laboratory data to identify important features of the fruit fly itself and the morphological relevance of male vs female parts; however the model is severely limited and cannot be expected to perform in this type of scenario without including the OOD data in training or making specialized processing for the anomalous cases.

Since the OOD set contains colors not heavily present in the training data, or contains colors that are used in both but in different contexts (like the red eyes vs the red raspberry surface), the actual results of segmentation do not have a very accurate grasp of the location of the fly. It is observed that the regions of segmentation which have tighter clusters of body identified pixels are predominately part of the fly, whereas pixels identified as body pixels that have diffuse clustering are generally not part of the fly. Other regions of contrast erroneously pick up edge and body pixels.

Still it is somewhat remarkable that the segmentation model is able to pick up on fly features from a data set with completely different camera, magnification, lighting, environmental and coloring conditions. It should be clear that color features are plausibly not the only factor influencing the confidence of network output.

3.2.8 An over-fit 224x224 experiment on SWD

One final example shows a model which spent too many epochs training (despite much augmentation). It decayed to about 76% binary classification accuracy after 300 epochs. The model's output is shown on a difficult sideways male example

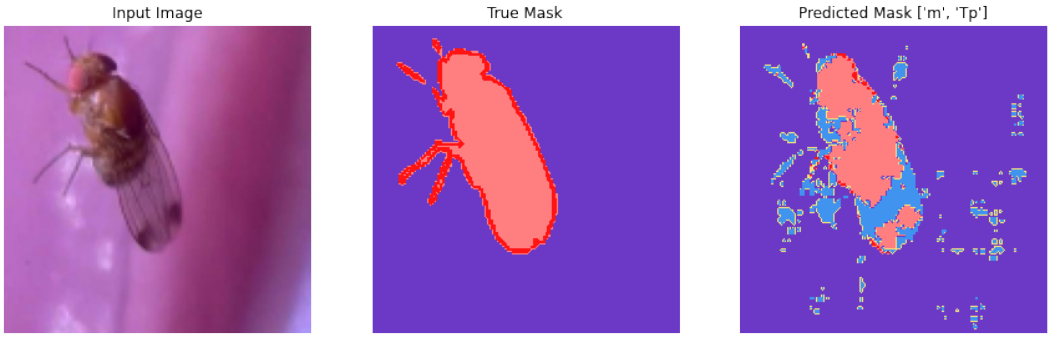

Figure 3.44: 128x128 prediction for a male SWD from the OOD data set. The background is of significantly different color than the training dataset, yet much of the prediction of body pixels contains fly. The spots on the wing are correctly distinguished as belonging to a male fly.

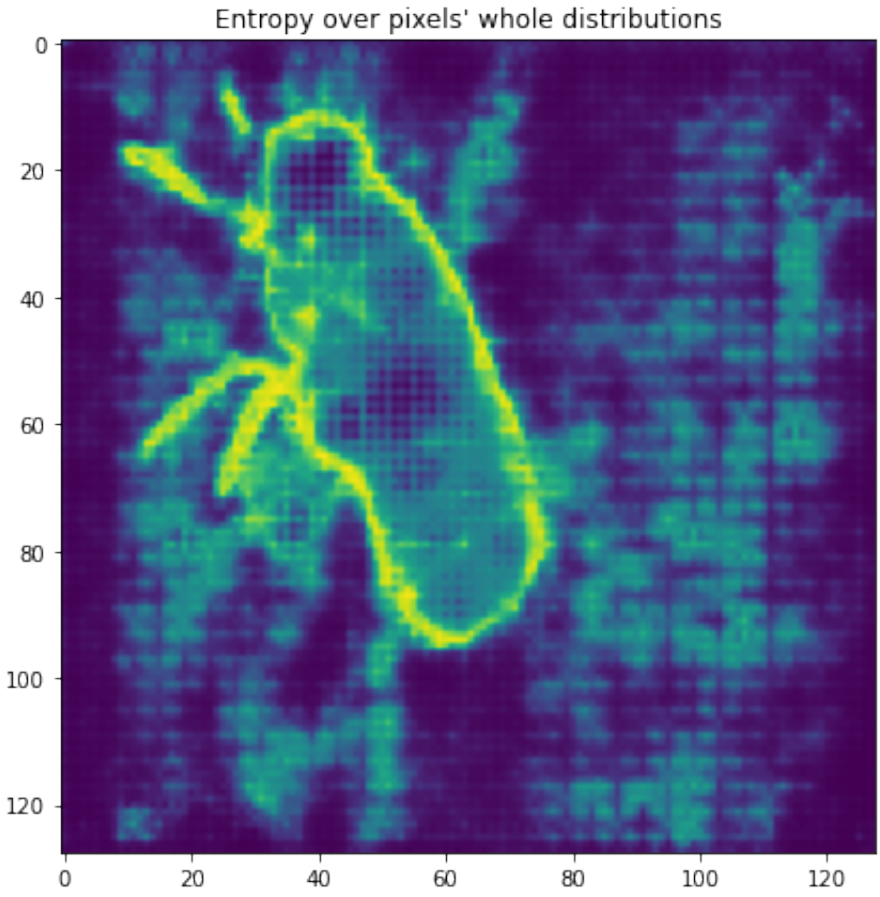

Figure 3.45: 128x128 pixel-wise entropy for a male SWD from the OOD data set. The entropy along the edges is as usual higher than the background entropy, though for this OOD example the background entropy is much higher than for the laboratory examples.

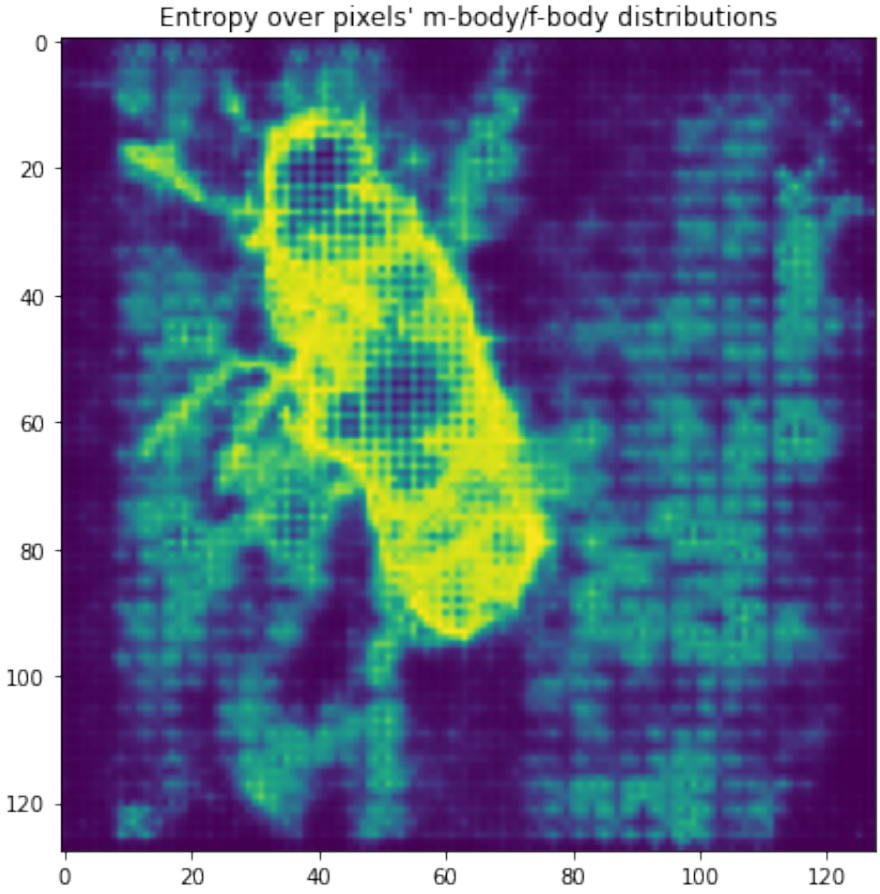

Figure 3.46: 128x128 pixel-wise entropy over only male and female body pixels for a male SWD from the OOD data set. The inner portion of the body still has strong contrast like it does in the laboratory examples, but the actual edges of the fly body still shine through as having higher entropy between male and female body pixels.

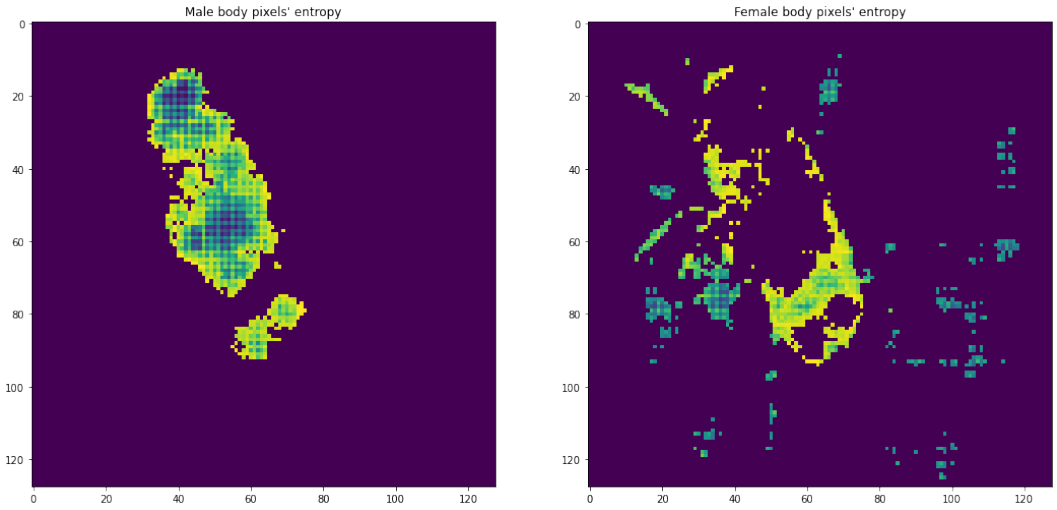

Figure 3.47: 128x128 pixel-wise entropy separated by classification to male and female body pixels. The male body pixels barely capture enough to contain the spots, while the female body pixels capture some of the body surrounding the spots and along the edges of the fly, in addition to the diffuse erroneous pixels around the picture.

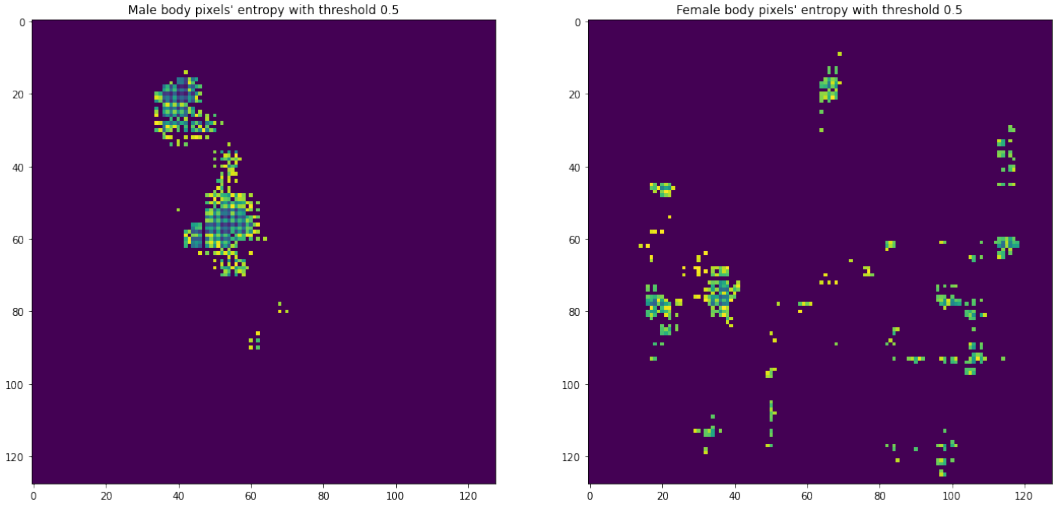

Figure 3.48: 128x128 pixel-wise entropy underneath the threshold 0.5 for male from the OOD set. The lower entropy portions of the male still include the spots, the bottom part as well as the head of the male. Whereas the female body pixels have largely diminished into misclassified background pixels. These background pixels are probably identifiable by the diffuse nature of their clustering in comparison to successful body pixel identifications.

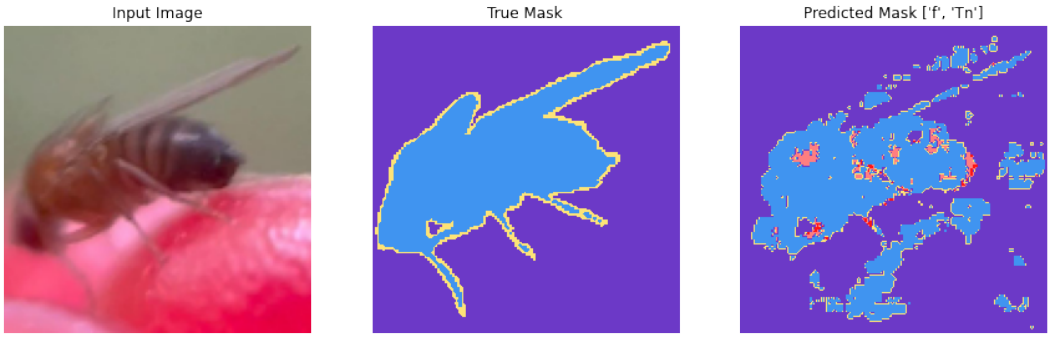

Figure 3.49: 128x128 prediction for sideways female from the OOD set. The female is successfully identified in this image, though the red shadow underneath along the raspberry attracts some misclassifications.

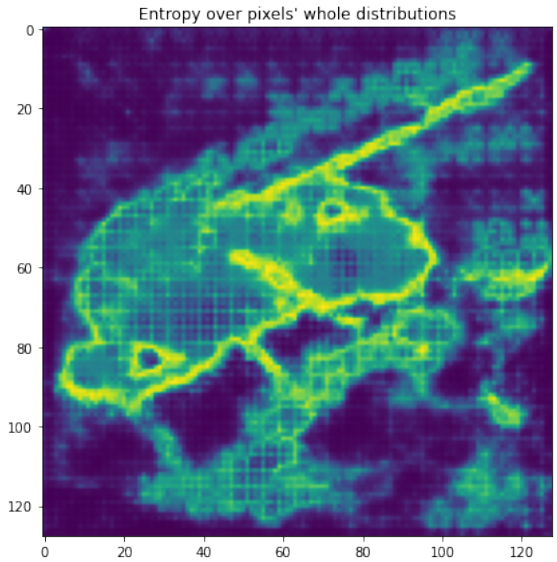

Figure 3.50: 128x128 pixel-wise entropy for sideways female from the OOD set. The edges of the female do continue to stand out as highly uncertain regions for the model.

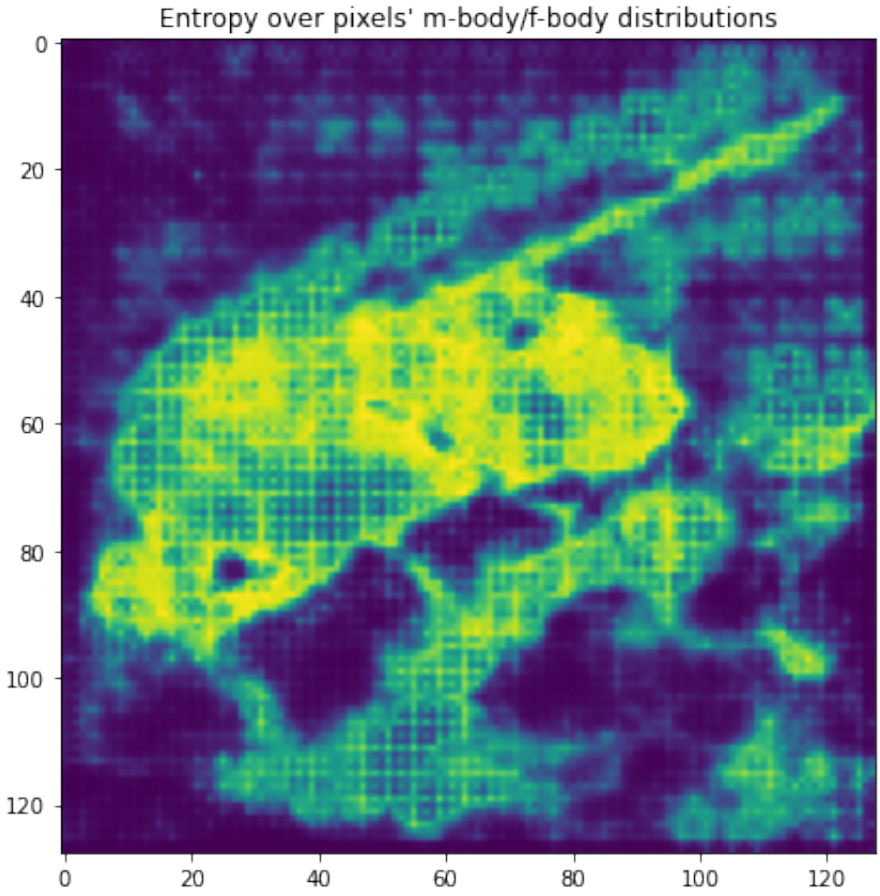

Figure 3.51: 128x128 pixel-wise entropy from male and female body pixel predictions for a sideways female in the OOD set. The actual body of the female has many of the highest parts of entropy within it. Though the high entropy of background features is still visible here at similar magnitudes to the lower entropy portions of the body.

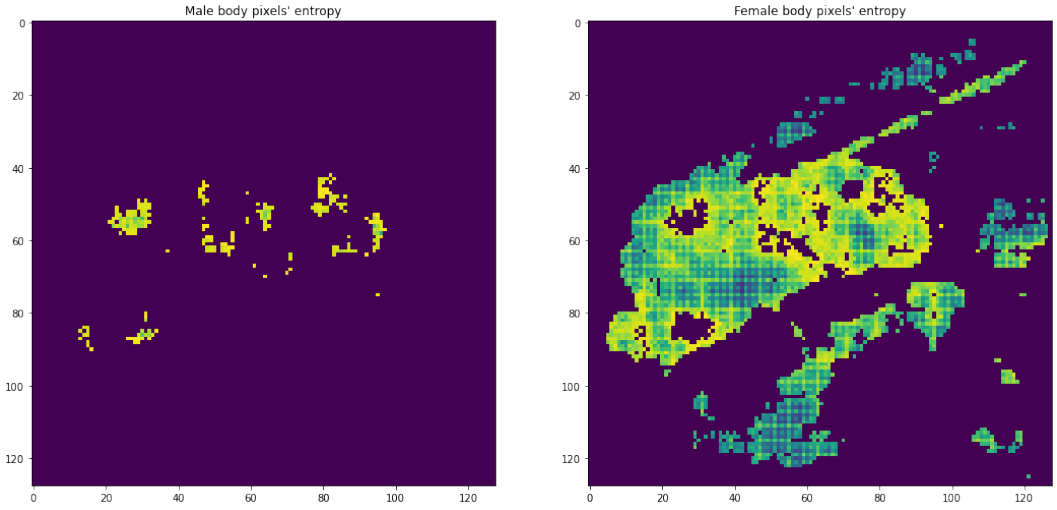

Figure 3.52: 128x128 entropy cut out between male and female predictions of body pixels for sideways female in the OOD set.

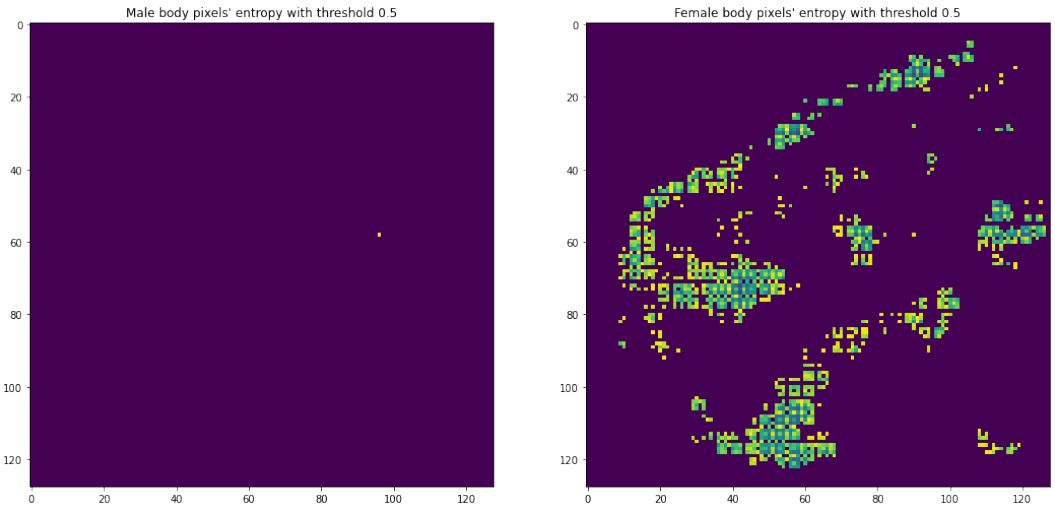

Figure 3.53: 128x128 entropy cut out between male and female predictions of body pixels and below threshold 0.5 for sideways female in the OOD set. While the lower entropy pixels capture some parts of the female body, her shadow still erroneously gets much of the attention.

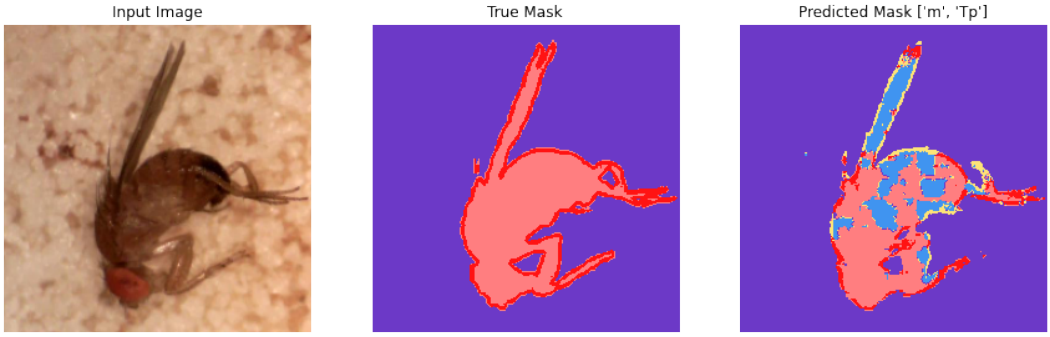

Figure 3.54: 224x224 prediction for sideways female. This example shows a situation where the wings are relatively uniform and appear to be female wings, so the model classifies them as female.

where the lighting conditions do not highlight the spots.

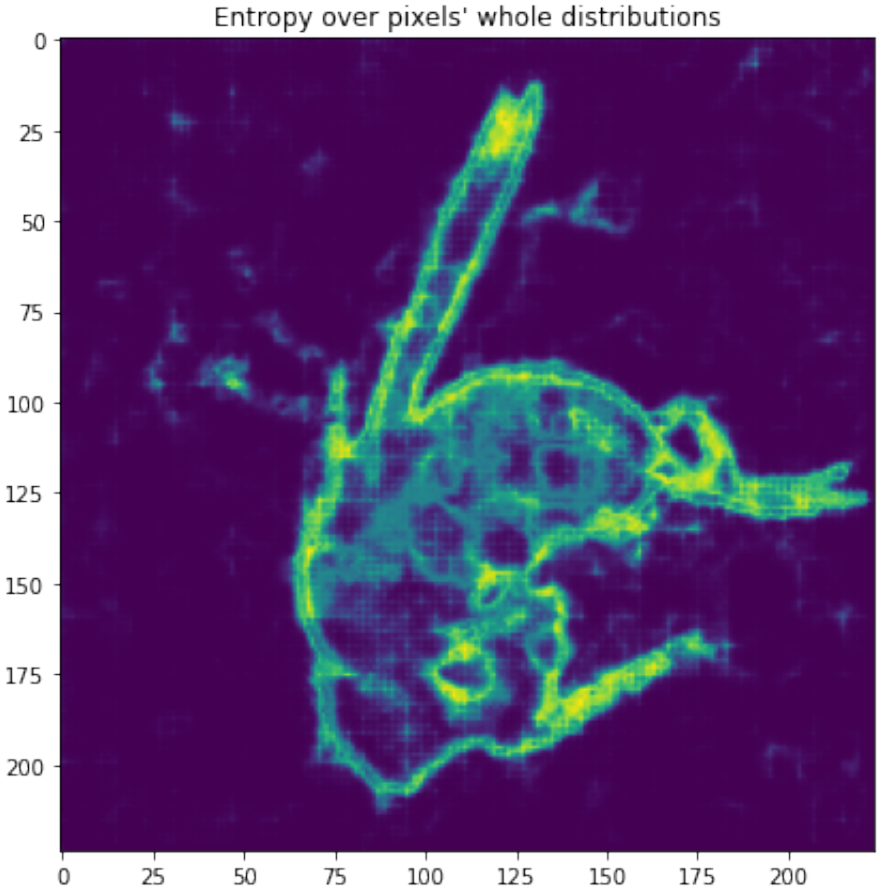

Figure 3.55: 224x224 pixel-wise entropy for a sideways male. The edge pixels are inhomogeneous and not as bright unlike most other examples.

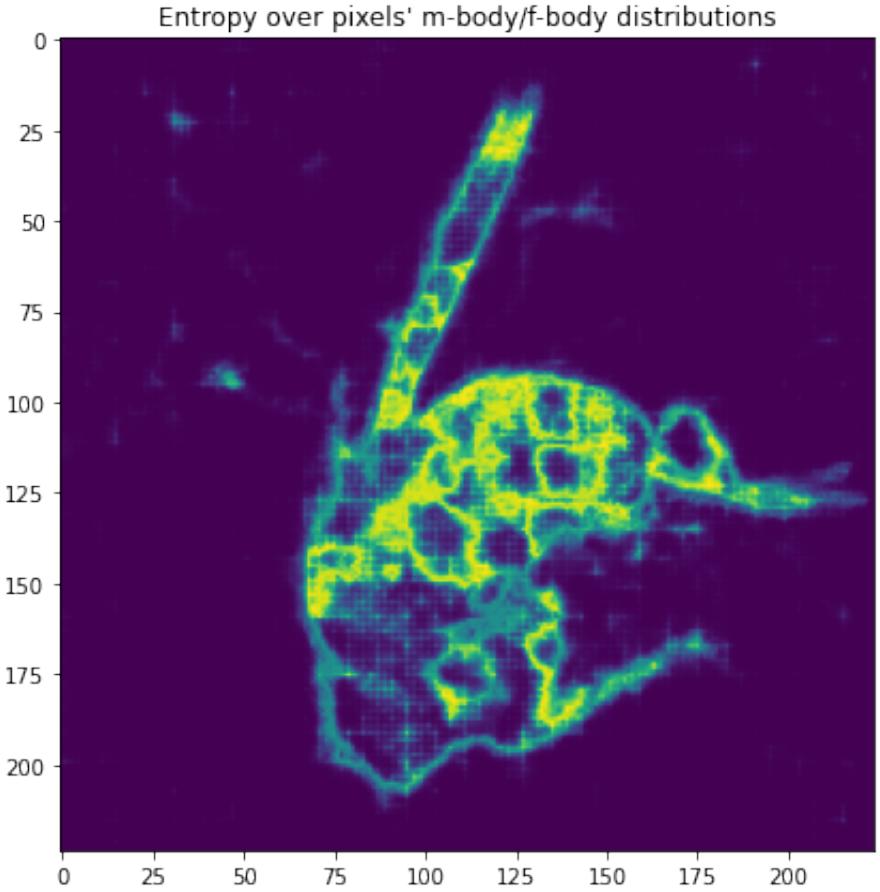

Figure 3.56: 224x224 pixel-wise entropy between only male and female body pixels for sideways male. Bright splotchy regions demonstrate the uncertainty between male and female classification. It is unclear from what cause the borders of these splotches originate, but possibly from the long over-fit time spent on this training run that allowed accuracy to decay.

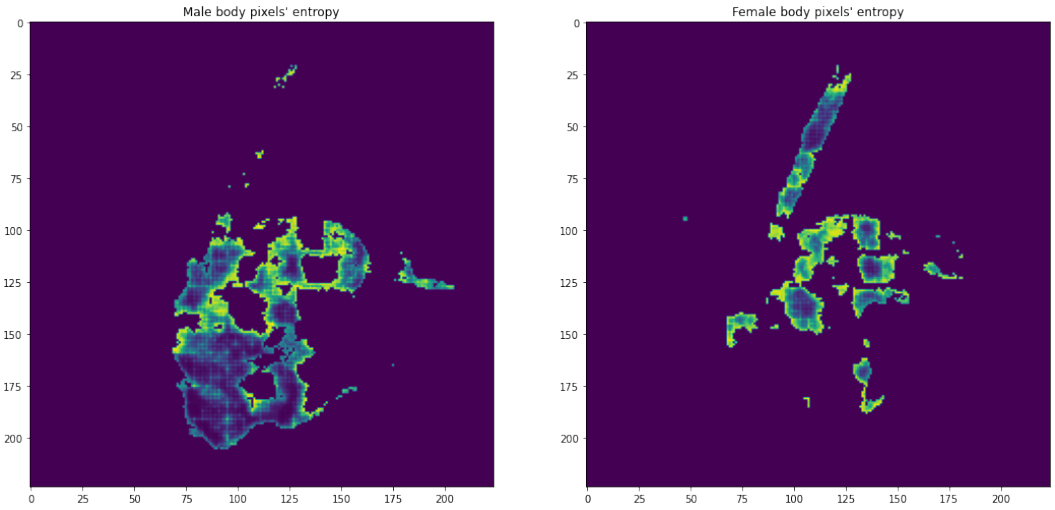

Figure 3.57: 224x224 pixel-wise entropy cut between male and female body pixels for a male on its side. The low entropy region along the wing where it is uniform and there is no spot visible that would normally identify a male indicates that this is a difficult example.

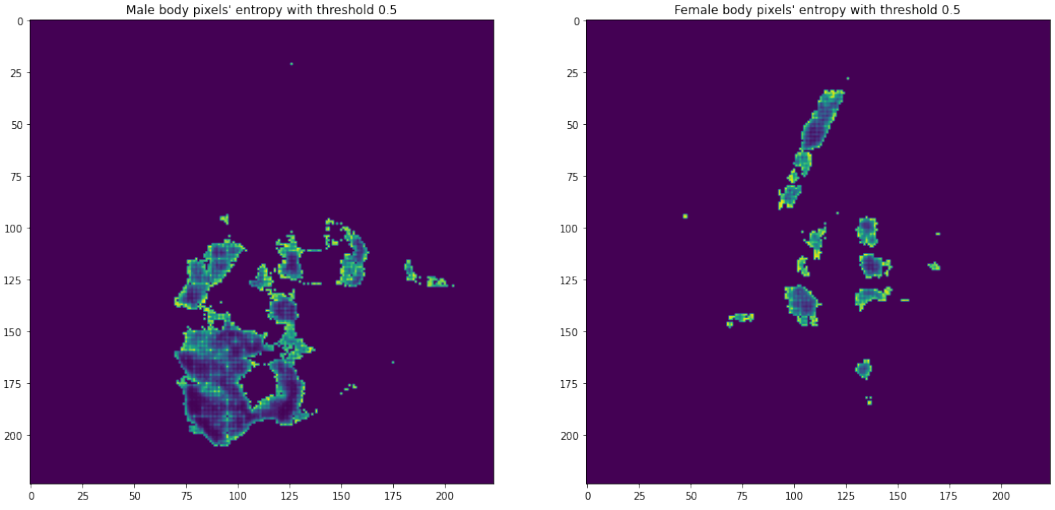

Figure 3.58: 224x224 pixel-wise entropy cut from male and female body pixels and under threshold 0.5 for a male on its side. The model has strong confidence that the wing belongs to a female, but the final prediction is still male from the dark bottom and the upper body and head pixels. Even with severe over-fitting the algorithm is able to get some successful classifications.

3.3 Discussion

A pre-trained MobileNetV2 + SWD-trained Pix2Pix proved to be a useful model to investigate which morphological features of SWD are accessible from pre-trained image encoding with a laboratory data set. OOD data was used to show that while high accuracy can be achieved in specialized situation such as restriction to laboratory data it does not necessarily carry over to alternative scenarios where the color contexts, backgrounds, lighting conditions and magnifications are different. Nonetheless the models were still able to exhibit some recognition of SWD in challenging situations including OOD data and difficult laboratory data situations where the most prominent feature (the spotted wing) is not visible. Binary accuracy as high as 85% was obtained for small resolutions, and with more exploration and improvements (or reductions) to augmentation the same models at higher resolutions should be able to score accuracy higher than 90% as indicated by results in the first chapter. The false negative rate was a nagging problem with the classification results and may possibly be reduced with relevant augmentation or collection of more data. This is not completely surprising because as was shown the certain angles of view of males can produce wings that look similar to female wings. Furthermore there are light gray flies in the data set which are difficult to identify since their sexual features may not have fully developed. Some of these are removed from the data set altogether and placed into a 'challenge' set which is tested and exhibits somewhat higher entropy. The segmentation algorithm still may have a better chance at identifying these flies than the amateur.

The question of whether this augmentation of spinning the fly on new backgrounds helps is still unanswered definitively, though in combination with other aug-

mentation it appeared to have effects on the confidence of the algorithms' outputs, including on OOD data, and especially with respect to the similar augmentations applied to the validation data upon output. This should be expected, though only useful if there is a situation that actually appears similar enough to the augmentations. At the levels of augmentation tested in this work the decoupling of segmented flies from their background had limited success and did not reliably improve accuracy. More work could be done to turn these segmentation examples into useful augmentation data.

Including the small exploration of effects of augmentation on training with segmentation data, this work also investigated pixel-wise entropy and found numerous instances where the difference in morphology of the fly between sexes revealed itself in the form of low entropy regions. Particularly, the spot on the male SWD wing was visible from predictions made by segmentation models. The dark regions on the on the males were key indicators which differentiated males and females. The algorithm did not place high confidence on the serrated ovipositor for the female in analysis of confidence, though usually a few pixels near the ovipositor were identified. At times the regions identified as highly confident included the legs, which may make one wonder whether the male sex combs played a role in identification. This is highly unlikely since segmentation was roughest along the legs and they were somewhat delicate to draw over. More work could be done to investigate whether the sex combs are visible in any of the images, potentially at the highest magnification 27x. The data set could be investigated and trimmed further to produce a more balanced set of prominent examples so that augmentation doesn't emphasize any misleading features, though much work has been done to comb the training data from these types of bad data including misclassifications by the author, which was a primary

source of inaccuracy at the start of the project where the spot on the wing was the only factor used to distinguish male from female.

The main contribution to this work was the collection of images of natural settings where the distribution of colors and complexity of the background varied significantly from the initial data used in training experiments. This out-of-distribution data was used to show that a model's classification accuracy is degraded when applied to observational settings significantly different from the settings under which the training data were observed. An augmentation environment was created to produce more training data and showed promise in improving the robustness of binary classification confidence under similar augmentations. A second contribution to this work was the use of pixel-wise entropy to offer a degree of interpretability of the localized confidence that a model has in its classification predictions when trained as a segmentation algorithm. Segmentation proved promising to extract a few of the prominent morphological features of the classification targets without being explicitly trained to output the highest confidence regions.

Future work could expand the data set and augmentation strategies to include observations of all insects in order to create a suite of data for any entomological studies suited to training algorithms across multiple tasks and distributions of data. Large scale pre-training that encompasses these data would see further improvements in accuracy. A rearing facility that could produce SWD in large numbers would also be capable of producing a suitable data set for the task of separation. The models used in this work would likely be able to filter out and pre-label a data set that could be easier and cheaper to prune down than it would be to build up from scratch by labeling data tediously, provided the rearing facility is able to produce enough flies and is equipped with equipment that can route the data to

the model. The segmentation algorithm may be more adept at detecting when the images produced are out of distribution intended for training.

Chapter 4

Conclusion

Quite a lot of material was explored in this work for the purpose of assessing feasibility of classifying between male and female Drosophila suzukii in SIT or other farm applications. The cost and difficulty of collecting large amounts of data is a serious hindrance to many industry applications and limits the time-span for which machine learning projects can be developed. We have seen that pre-trained models are wonderful additions to the community which reduce the amount of data required to perform image classification tasks.

4.1 Discussion of the main objectives and results

Reflecting back on the objectives set out in this work, we can ask the question: to what extent are available machine learning algorithms effective at classifying SWD, and how much data is required to eclipse the 90% accuracy threshold (and further)? With less than 1000 examples we are able to classify SWD to accuracy above 90% with both newer and slightly aged pre-trained models with minimal architectural design required on top of them. It is clear that these models greatly increase accessibility of intelligent technology, which we have seen applied in this

work to situations which will be useful for SIT and other ecological and agricultural monitoring situations. Pre-trained models show fantastic transfer learning capability for identification of sex of fruit flies.

Of the available models CLIP, particularly the ViT-16 variant, performs the best, and with the addition of the text encoder this model also has many zero-shot flexibility applications which were not explored in this work. ImageNet is one diverse resource that brings tremendous applicability to the domain of segmenting and classifying SWD, and while the models did not perform better than CLIP, they still were able to eclipse 90% accuracy on the main data set. These models, like MobileNetV2, served as a benchmark for classification tasks that might be geared towards usage on user-devices rather than on the cloud. Although the CLIP models are able to be computed in reasonable times on ordinary devices as well, they would also be suitable to be evaluated remotely on the cloud. The collective results of the models demonstrate the steady increase in performance of models that has been observed this decade.

Robustness to out-of-distribution data from one set of environmental conditions to another will be a desirable trait for classification models. For the purposes of this work we have seen that expansions of the data set into OOD settings may be required in order to use the same classification model on applications outside of SIT, such as observing SWD in nature or on the farm, or in laboratory or rearing facility settings which use environments that appear different than the data originally collected. For classification applications to be developed the specific use-case setting should be clearly defined and tuned to the situation where it will be applied; however it is possible that further improvements in models and consolidation of data sets into canonical examples relevant to the environmental settings in the expected

use-case will allow for more flexibility in the task. For this purpose the development of an augmentation environment was used to investigate both the models and the data set. Preliminary observations showed that augmentation can increase the confidence of the machine learning algorithm even on data which it has not been trained on. The augmentation environment must be created so that the alterations to images are realistic enough to be useful. While further work is needed to investigate if this augmentation environment will be effective at extending applications into OOD domains or reducing the overall amount of data required for training, the advancements leading to robust models like CLIP seem to outweigh the impact of hand-crafted augmentation environments. The augmentation environment did prove useful in detecting instances of false negative samples in the training data when observing the learning curve of the confusion matrix, since excessive augmentation of false negative examples will lead to an imbalance in the confusion matrix which gradually reduces the recall as training progresses over many epochs. This would of course be observable for false positives as well, though for SWD it is more difficult to make false positives than false negatives since the distinguishing factor of the spot on the male's wing is present in most if not all positive examples. Despite numerous examples in the validation data set containing flies whose side-ways orientation limited the visibility of the dark spot, the algorithm was still able to successfully classify most of these. Upon recovering the small number of false negatives and removing them from the training data it was observed that the accuracy of the algorithm was increased.

We have also seen how image segmentation can be used to further investigate the data set used for training. Much concern in the machine learning community has been focused on explainability and interpretability of the decisions

that any instance of an algorithm makes. While image segmentation in and of itself provides fine-grained classification to images, investigation of the confidence of these predictions can lead to morphological factors in the data set which are relevant for classification but were not directly targeted in the training data. These observations were only made using the MobileNetV2 pre-trained model as a base for image segmentation and should be further explored with other models and data sets before coming to a definitive conclusion; however, research efforts along these lines would contribute to improving human confidence in interpreting results from machine learning algorithms, and indeed many deep explorations in this topic have already been made independently within the short time-span that this project was carried out, such as those which have improved on the LIME algorithm which were used initially to assess interpretability of the binary classification results. In the case of this project image segmentation provided more information about the usefulness of the data set than did the implementation of the LIME algorithm for binary classification, but it is possible that more fine-tuned or updated usages of LIME could be equally useful.

Future pre-trained models with diverse sets of data which include biological data may further improve the accuracy and robustness. Combination of fitting data sets to models that use pre-trained image encoding and language prompts will likely continue to be used to great effect. The amount of data required for these classification tasks, excluding the vast amount of data used for pre-training, will gradually (or suddenly) be reduced to zero, eventually. Future work in IPM will have the opportunity to use other methods in the machine learning realm.

The main contribution of this work is the collection of a data set and applied image classification algorithms which will be useful for implementing SIT

to reduce and prevent the destruction of crops of stone-fruits by the non-native Drosophila suzukii species, which looks to solve a problem costing upwards of $1 Billion dollars per year when accounting for inflation and global crop production. Another key contribution includes the collection of an OOD data set that assesses the ability to extend beyond the SIT application and into monitoring for SWD in traps on farms or out in the field, as monitoring the effects of any SIT operation on the population of fruit flies will be essential in determining the effectiveness of the technique. Finally, the demonstration that existing community models have been very effective at classifying SWD provides a modest snapshot of some of the capabilities of machine learning technology developed in the last decade. Models developed and delivered by the community will be essential for applications such as SIT which enhance and protect the food security of our society and further development should be encouraged.

4.2 Future work and outlook

The contributions discussed in the previous section may be of value to practitioners in academia and industry. The results of developing a classification algorithm are easily transferred to the cloud and evaluated with speed much greater than that of devices like personal computers and cell phones. A photograph taken from a trap in the field or an imaging system in a rearing facility could be quickly uploaded and used with any one of the models trained to separate male from female SWD. These pre-trained models are used as a demonstration in this work to show their effectiveness at an industry application that has potential to be highly valuable. Their applicability could be eclipsed by a new, larger data set collected in real-time

at a rearing facility or a collection of SWD traps, potentially using models applied in this work to automatically label the newer data as it passes through the imaging devices. Usage of these models for the purpose of initially labeling data could save researchers and experts much time in sifting through and tediously labeling the data by hand. If there is an error rate of 5%, this will potentially save 95% of the time spent labeling and free up the researchers to move on to their next task. It was observed through the augmentation experiments that different instances of the same fly may also be classified as accurate in the face of an initial false identification when passed through the augmentation algorithms. A threshold on the confidence of the algorithm in order to make the decision to include or exclude a positive identification could exclude flies for which the algorithm has a high false negative rate, though low entropy false negative and positive examples should be carefully subjected to scrutiny. Re-running the set of leftover flies through the data set under different experimental conditions such as lighting and magnification alterations could further improve the concentration of the collection of positive samples.

The computer science community continues to develop artificial intelligence algorithms at a rapid pace, and for the purposes of this work their algorithms have enabled collaboration between scientists as diverse as physicists and biologists to target food security in a way that will ultimately, it is hoped, be beneficial to society. While much of deep learning in the previous decade was characterized by requiring huge amounts of data for training, this decade is seeing advancements that lower the data threshold needed to make machine learning projects successful. Scaling laws in machine learning will mean continued improvement which has been leading to zero-shot classification results in many domains (e.g. with no additional training data), which could potentially make many job skills obsolete (including program-

mers), freeing much of the labor force to change the nature of their interaction with industry and potentially transferring the value obtained from automation directly to benefit society. While the future is uncertain, advances from technologies developed in the previous decade have already made impacts in numerous domains and will likely continue to quietly enhance the products we already use and lead to new revolutionary products as well.

Bibliography

[1] Zhu M. Zhmoginov A. Chen L-C. Sandler M., Howard A. Movilenetv2: Inverted residuals and linear bottlenecks. *arXiv preprint arXiv:1801.04381v4*, 2019.

[2] Hallacy C Ramesh A Goh G-Agarwal S Sastry G Askell A Clark J Krueger G Sutskever I Radford A, Kim JW. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020v1*, 2021.

[3] Kolesnikov A Wessenborn D Zhai X-Unterthiner T Dehghani M Minderer M Heigold G Gelly S et al. Dosovitskiy A, Beyer L. An image is worth 16x16 words: Transformers for image recognition at scale. *Pest Manag Sci*, 2017.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[5] Richard Zhang. Making convolutional networks shift-invariant again, 2019.

[6] Le Q Tan M. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[7] Rachael E.; Zalom Frank G. Bolda, Mark P.; Goodhue. Spotted wing drosophila: Potential economic impact of newly established pest. *Agricultural and Resource Economics Update (ARE Update)*, 13(3):5–8, 2010.

[8] Toral; Finch Deborah M.; Miniat-Chelcy Ford; Hayes Deborah C.; Lopez Vanessa M. Poland, Therese M.; Patel-Weynand. *Invasive Species in Forests and Rangelands of the United States*. Springer International Publishing, 2021.

[9] Usda awards $6.7 million to stifle spotted wing drosophila. *Growing Produce*, 2015.

[10] Mark P.; Goodhue Rachael E.; Dreves-Amy J.; Lee Jana; Bruck Denny J.; Walton Vaughn M.; O'Neal Sally D.; Zalom Frank G Walsh, Douglas B.; Bolda. Drosophila suzukii (diptera: Drosophilidae): invasive pest of ripening soft fruit expanding its geographic range and damage potential. *Journal of Integrated Pest Management*, 2(1):G1–G7, 2011.

[11] Michelle T Fountain, Amir Badiee, Sebastian Hemer, Alvaro Delgado, Michael Mangan, Colin Dowding, Frederick Davis, and Simon Pearson. The use of light spectrum blocking films to reduce populations of drosophila suzukii matsumura in fruit crops. *Scientific Reports*, 10(1):1–12, 2020.

[12] Sterile insect technique. *https://www.iaea.org/topics/sterile-insect-technique*.

[13] Rodriguez-Saona C Short BD Kirkpatrick DM Loeb GM Aflitto NC Wiman N Andrews H Drummond FA Fanning PD Ballman E Johnson B Beal DJ Beers EH Burrack HJ Isaacs R Perkins J Liburd OE Lambert AR Walton VM Harris ET Mermer S Polk D Wallingford AK Adhikari R Sial AA Panthi B, Cloonan KR. Using red panel traps to detect spotted-wing drosophila and its infestation in us berry and cherry crops. *J Econ Entomol*, 2022. 115(6):1995-2003. doi: 10.1093/jee/toac134. PMID: 36209398.

[14] WHO. Global vector control response 2017-2030. *WHO*, 47, 2017.

[15] Martin Hauser. A historic account of the invasion of drosophila suzukii (matsumura)(diptera: Drosophilidae) in the continental united states, with remarks on their identification. *Pest management science*, 67(11):1352–1357, 2011.

[16] Jana C Lee, Denny J Bruck, Amy J Dreves, Claudio Ioriatti, Heidrun Vogt, and Peter Baufeld. In focus: spotted wing drosophila, drosophila suzukii, across perspectives. *Pest management science*, 67(11):1349–1351, 2011.

[17] Eschen R. Kenis M, Tonina L. Non-crop plants used as hosts by drosophila suzukii in europe. *J Pest Sci*, 89:735–748, 2016.

[18] Walton VM. Klick J, Yang WQ. Distribution and activity of drosophila suzukii in cultivated raspberry and surrounding vegetation. *J Appl Entomol*, 140:37–46, 2015.

[19] Bolda M Goodhue RE Williams JC Zalom FG Farnsworth D, Hamby KA. Economic analysis of revenue losses and control costs associated with the spotted wing drosophila, drosophila suzukii (matsumura), in the california raspberry industry. *arXiv preprint arXiv:2010.11929*, 2020. 73(6):1083-1090. doi: 10.1002/ps.4497. Epub 2017 Jan 26. PMID: 27943618.

[20] Jones R Gilbert D Mckemey AR Slade G Fountain MT Homem RA, Mateos-Fierro Z. Field suppression of spotted wing drosophila (swd) (drosophila suzukii matsumura) using the sterile insect technique (sit). *Insects*, 2022. Mar 26;13(4):328. doi: 10.3390/insects13040328. PMID: 35447770; PMCID: PMC9031279.

[21] Fournier F Martel V Vreysen M Cáceres C Firlej A Lanouette G, Brodeur J. The sterile insect technique for the management of the spotted wing drosophila,

drosophila suzukii: Establishing the optimum irradiation dose. *PLoS One*, 2017. 12(9):e0180821. doi: 10.1371/journal.pone.0180821. PMID: 28957331; PMCID: PMC5619704.

[22] Mouton L Stauffer C Bourtzis K Nikolouli K, Sassù F. Combining sterile and incompatible insect techniques for the population suppression of drosophila suzukii. *J Pest Sci*, 2004. 93(2):647-661. doi: 10.1007/s10340-020-01199-6. Epub 2020 Jan 29. PMID: 32132880; PMCID: PMC7028798.

[23] Belikoff EJ Berger A Griffith EH Scott MJ Li F, Yamamoto A. A conditional female lethal system for genetic suppression of the global fruit crop pest drosophila suzukii. *Pest Manag Sci*, 2021. 77(11):4915-4922. doi: 10.1002/ps.6530. Epub 2021 Jul 2. PMID: 34169646.

[24] Stockton D Lee J Avosani S Abrieux A Anfora G Beers E Biondi A Burrack H Cha D Chiu JC Choi MY Cloonan K Crava CM Daane KM Dalton DT Diepenbrock L Fanning P Ganjisaffar F Gómez MI Gut L Grassi A Hamby K Hoelmer KA Ioriatti C Isaacs R Klick J Kraft L Loeb G Rossi-Stacconi MV Nieri R Pfab F Puppato S Rendon D Renkema J Rodriguez-Saona C Rogers M Sassù F Schöneberg T Scott MJ Seagraves M Sial A Van Timmeren S Wallingford A Wang X Yeh DA Zalom FG Walton VM Tait G, Mermer S. Drosophila suzukii (diptera: Drosophilidae): A decade of research towards a sustainable integrated pest management program. *J Econ Entomol*, 2021. 114(5):1950-1974. doi: 10.1093/jee/toab158. PMID: 34516634.

[25] Vreysen MJB Bourtzis K. Sterile insect technique (sit) and its applications.

Insects, 2021. 12(7):638. doi: 10.3390/insects12070638. PMID: 34357298; PMCID: PMC8304793.

[26] Imagenet overview. https://image-net.org/about.php.

[27] Su H Krause J Satheesh S Ma S Huang Z Karpathy A Khosla A Bernstein M Berg A Fei-fei L Russakovsky O, Deng J. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 2015. 115(3):211–252.

[28] Lee K Toutanova K Devlin J, Chang MW. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2020.

[29] Henighan T Brown TB Chess B Child R Gray S Radford A Wu J Amodei D. Kaplan J, McCandlish S. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[30] Ryder N Subbiah M Kaplan J Dhariwal P Neelakantan A Shyam P Sastry G Askell A et al. Brown TB, Mann B. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[31] OpenAI. Gpt-4 technical report, 2023.

[32] Methods of insect image capture and classification: A systematic literature review. *Smart Agricultural Technology*, 1:100023, 2021.

[33] Sana Tariq, Ayesha Hakim, Awais Ahmad Siddiqi, and Muhammad Owais. An image dataset of fruitfly species (bactrocera zonata and bactrocera dorsalis) and automated species classification through object detection. *Data in Brief*, 43:108366, 2022.

[34] Chumchuen K Samung Y Sriwichai P Phatthamolrat N Tongloy T Jaksukam K Chuwongin S Boonsang S Kittichai V, Pengsakul T. Deep learning approaches for challenging species and gender identification of mosquito vectors. *Nature: scientific reports*, 11(4838), 2021.

[35] Yingqiong Peng, Muxin Liao, Hong Deng, Ling Ao, Yuxia Song, Weiji Huang, and Jing Hua. Cnn–svm: a classification method for fruit fly image with the complex background. *IET Cyber-Physical Systems: Theory & Applications*, 5(2):181–185, 2020.

[36] Gonzalo I González-López, G Valenzuela-Carrasco, Edmundo Toledo-Mesa, Maritza Juárez-Durán, Horacio Tapia-McClung, and Diana Pérez-Staples. Determination of the physiological age in two tephritid fruit fly species using artificial intelligence. *Journal of Economic Entomology*, 115(5):1513–1520, 2022.

[37] Fábio Augusto Faria, Paula Perre, Roberto A Zucchi, Leonardo Ré Jorge, TM Lewinsohn, Anderson Rocha, and R da S Torres. Automatic identification of fruit flies (diptera: Tephritidae). *Journal of Visual Communication and Image Representation*, 25(7):1516–1527, 2014.

[38] Matheus Macedo Leonardo, Tiago J. Carvalho, Edmar Rezende, Roberto Zucchi, and Fabio Augusto Faria. Deep feature-based classifiers for fruit fly identification (diptera: Tephritidae). In *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 41–47, 2018.

[39] Zisserman A Simonyan K. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. 2018.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[42] Wu J. Child R. Luan D. Amodei D. Radford, A. and Sutskever I . Language models are unsupervised multitask learners. 2019.

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[44] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[45] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[46] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks, 2018.

[47] Lisa J Amstutz. *Invasive Species*. Abdo Publishing, 2018.

[48] Lynell Tanigoshi Jimmy Klick Joseph Kleiber Joe DeFrancesco Beverly Gerdeman Hollis Spitler Denny J Bruck, Mark Bolda. Laboratory and field comparisons of insecticides to reduce infestation of drosophila suzukii in berry crops. *Pest Manag Sci*, 67(11):1375–1385, 2011.

[49] Hugh J. Colautti, Robert I.; MacIsaac. A neutral terminology to define 'invasive' species. *Diversity and Distributions*, 10(2):135–141, 2004.

[50] Lee JC; Dreves AJ; Cave AM et al. Infestation of wild and ornamental noncrop fruits by drosophila suzukii, (diptera: Drosophilidae). *Ann Entomol Soc Am*, 108:117–129, 2015.

[51] D. Pimentel, D.; Zuniga; Morrison. Update on the environmental and economic costs associated with alien-invasive species in the united states. *Ecological Economics*, 52(3):273–288, 2005.

[52] Rodrigo Lasa, Eduardo Tadeo, Ricardo A. Toledo-Hérnandez, Lino Carmona, Itzel Lima, and Trevor Williams. Improved capture of drosophila suzukii by a trap baited with two attractants in the same device. *PLOS ONE*, 12:1–19, 11 2017.

[53] Executive order 13112 – 1. definitions. Archived from the original June 25, 2021: https://web.archive.org/web/20210625075018/https://www.invasivespeciesinfo.gov/executive-order-13112-section-1-definitions.

Appendix A

Supplementary material for binary classification with pre-trained networks

A.1 5-fold cross-validation of binary classification without augmentation

The main goal within the scope of these experiments was to be able to produce an effective algorithm to identify the sex of SWD given a relatively small sized data set. Augmentation was necessary to improve the accuracy significantly above the 90% mark, though it is remarkable that the CLIP models were able to achieve close to 90% accuracy on a data set of less than 800 low-resolution images (224x224). These results are comparable to and even better than results from other experiments which used higher resolutions or trained their models from scratch. This is a testament to the massive pre-training carried out for the CLIP models and points to the impressive performance gains achieved by machine learning researchers over the last decade.

Accuracy				
Pretrained model	SGD	LinSVC	PolySVC	LogReg
ViT-B/32	0.8875 ±0.0208	0.8935 ±0.0056	0.8120 ±0.0244	0.8889 ±0.0265
ViT-B/16	0.8980 ±0.0434	0.8874 ±0.0278	0.8507 ±0.0171	0.9054 ±0.0152
RN50x16	0.8832 ±0.0162	0.8743 ±0.0295	0.8269 ±0.0297	0.8803 ±0.0200
RN50x4	0.8713 ±0.0121	0.8670 ±0.0261	0.8017 ±0.0253	0.8624 ±0.0071
RN101	0.8729 ±0.0318	0.8608 ±0.0323	0.8463 ±0.0208	0.8579 ±0.0323
RN50	0.8875 ±0.0297	0.9037 ±0.0187	0.8210 ±0.0173	0.8949 ±0.0305
VGG16	0.7911 ±0.0548	0.8417 ±0.0340	0.5176 ±0.0331	0.8374 ±0.0264
VGG19	0.7901 ±0.0386	0.8328 ±0.0408	0.5132 ±0.0321	0.8475 ±0.0434
MobileNetV2	0.7708 ±0.0379	0.8077 ±0.0481	0.5307 ±0.0682	0.8331 ±0.0408

Table A.1: These are the results of training classifiers for SWD without augmentation. The RN50 and ViT-B/16 models perform the best, eclipsing 90% accuracy. Without augmentation the CLIP models perform better than the MobileNetV2 and VGG models. As many as 10 false negative classifications contaminated the data set for this result, which may be 1-2% lower in accuracy across all models due to the contamination.

A.2 OOD evaluation of binary classification results

One of the hopes for this algorithm was that it would be applicable to domains that extend beyond the factory or laboratory setting to natural environments or testing environments that exist on the farms themselves.

The OOD data set introduced in chapter 3 was not large enough or diverse enough to use for training data, and therefore the evaluation of this data in aggregate may not be a strong indication for a robust algorithm; however, it does give some information about the effectiveness of algorithms on data which they were not intended to be trained. It was observed that the ResNet models performed better for

CLIP on OOD data than did the ViT models. This is somewhat surprising because the ViT models were more accurate on data for which they were fine-tuned (rather than the OOD data). This may simply be due to an accuracy-robustness trade-off, or it may be that the ResNet models really are more robust. Further investigation of the robustness of the CLIP models is warranted. MobileNetV2 also performed admirably on the OOD data in comparison to its VGG counter-parts and added to its appeal for use during segmentation in addition to the fact that it is faster to evaluate.

Accuracy on OOD data				
Pretrained model	SGD	LinSVC	PolySVC	LogReg
ViT-B/32	48.684	46.053	60.526	46.053
ViT-B/16	52.632	50.000	57.895	50.000
RN50x16	61.842	59.211	61.842	64.474
RN50x4	63.158	50.000	63.158	53.947
RN101	65.789	71.053	69.737	69.737
RN50	56.579	57.895	56.579	55.263
VGG16	59.211	53.947	44.737	52.632
VGG19	48.684	61.842	64.474	61.842
MobileNetV2	64.474	68.421	51.316	68.421

Table A.2: A table showing accuracy on OOD data for binary classification with different pre-trained models used with augmentation. Accuracy should not be expected to be high for this data since it differs from the laboratory data substantially; however the ResNet models appear to perform better on the OOD data than the ViT models despite having lower accuracy than the ViT models on in-distribution data.

A.3 LIME boundaries for selected examples

A diagnostic used for evaluating results of a model was the so-called LIME algorithm. The major advantage of this algorithm is that it is model-agnostic – it treats any model like a black box and attempts to reconstruct the regions of the image which are responsible for a negative or positive classification.

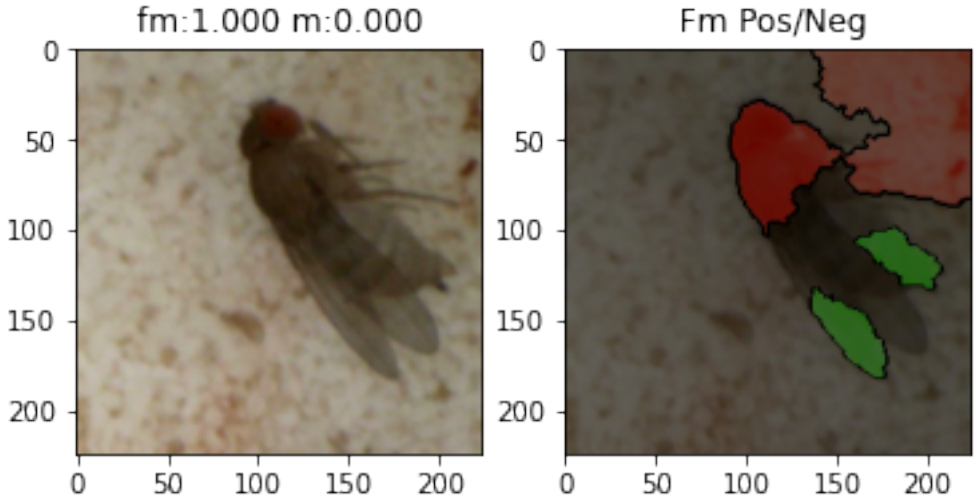

Figure A.1: Left: An in-distribution validation example correctly classified as female with prediction probability equal to one using the ViT-B/16 model. Right: An example of LIME boundaries calculated for the same image, green regions are identified as female and red regions are identified as male.

The results of using this algorithm were highly variable and it was not as useful as was hoped for; however it did produce some results potentially consistent with the pixel-wise entropy images displayed in chapter 3. In particular LIME often identified the wings, or regions near the wings as well as regions near the serrated ovipositor of the females and darkened inward facing bottoms of males. The heads also were occasionally a region of interest, even though this author did not see the head region as an especially useful region for classification. LIME also frequently identified regions of the background, often bordering the fly, as places of positive or negative identification. This was part of the motivation for expanding augmentation of training data by using segmentation since it was expected that over-fitting to the background was occurring. Therefore an attempt was made to decouple the backgrounds from the fruit flies in the original images.

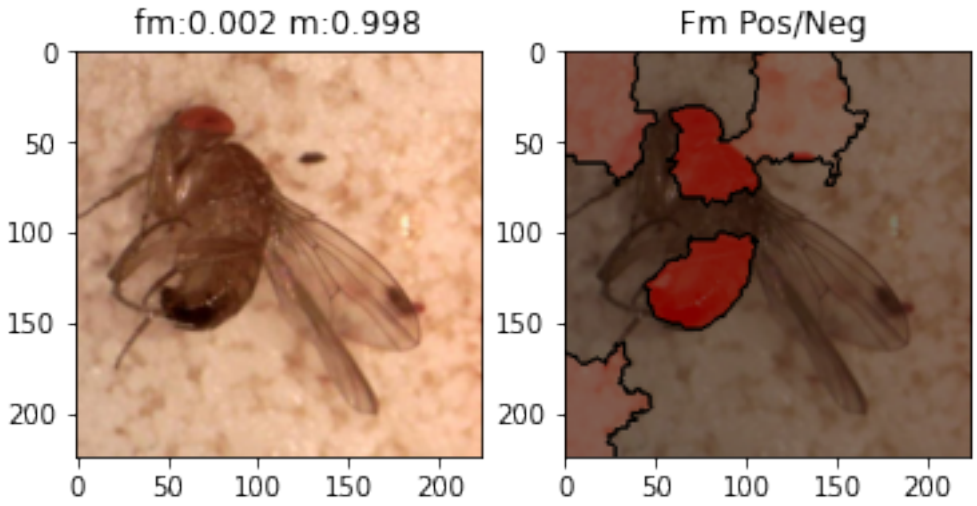

Figure A.2: Left: An in-distribution validation example correctly classified as male using the ViT-B/16 model. Right: An example of LIME boundaries calculated for the same image, green regions are identified as female and red regions are identified as male; there are only red regions for this LIME calculation.

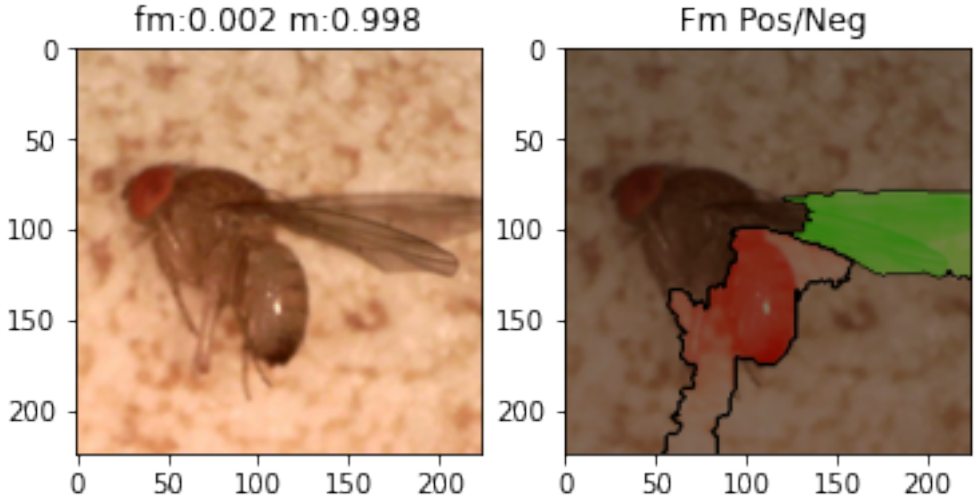

Figure A.3: Left: An in-distribution validation example correctly classified as male using the ViT-B/16 model. Right: An example of LIME boundaries calculated for the same image, green regions are identified as female and red regions are identified as male; this example was initially incorrectly labeled as female by the data labeler; however the algorithm predicts with high confidence that it is male. This example is considered a difficult example because the wings do not clearly show the spot at this camera angle, and the lack of serrated ovipositor may not be completely obvious.

Appendix B

Supplementary material for image segmentation with pre-trained networks

B.1 Augmentation strategies used during segmentation

One of the goals of creating the segmentation data set was to improve the efficiency of training data in various ways. This should be clear that the segmentation data contains pixel-wise information produced by the data labeler, whereas the binary labeling used for the results in chapter 2 contained less information. While the segmentation algorithms tested did perform with higher binary classification accuracy (comparing to the binary classification which used MobileNetV2), another expectation was that the augmentation could be further improved as well.

In particular it would seem that cutting the body of the fly and pasting a rotated version onto various backgrounds should be useful for data augmentation – more useful than merely rotating the image with the default TensorFlow or PyTorch

augmentations.

This proved more difficult than expected, as the segmentation algorithms performed with better binary accuracy when using less of this augmentation rather than more. The suspicion is that augmentations using this strategy may be good, but the magnitude of the change to the data set may be too large to be observed during this project.

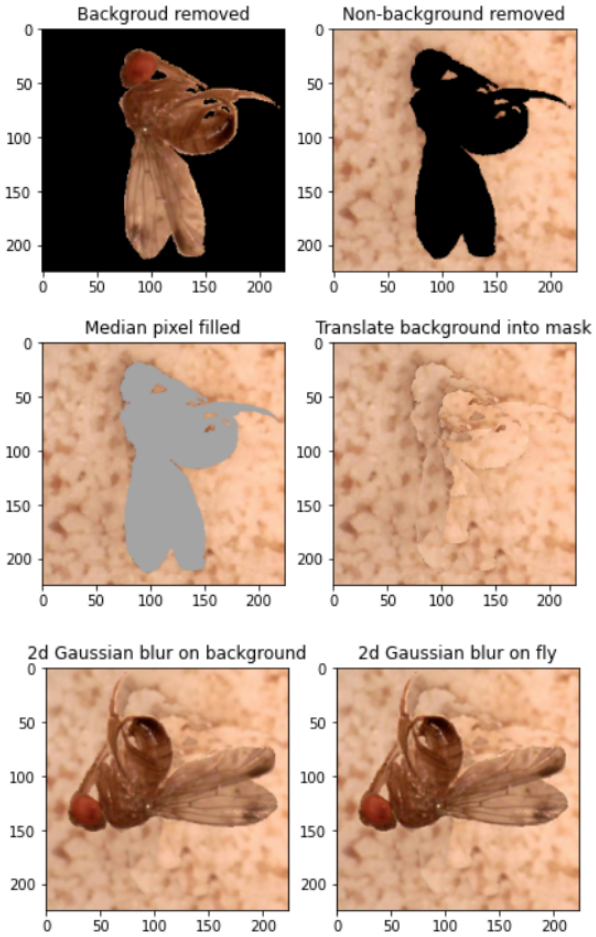

Figure B.1: The segmentation training data allowed for additional augmentation to be developed. A mask containing the body of the fly as well as a separate mask for the background were created so that the fly could be decoupled from the background during training. The empty fly region was filled with the median value of the background image to patch over any parts the translated background might have missed. For the actual training, backgrounds were randomly selected instead of using the same background from the original image.

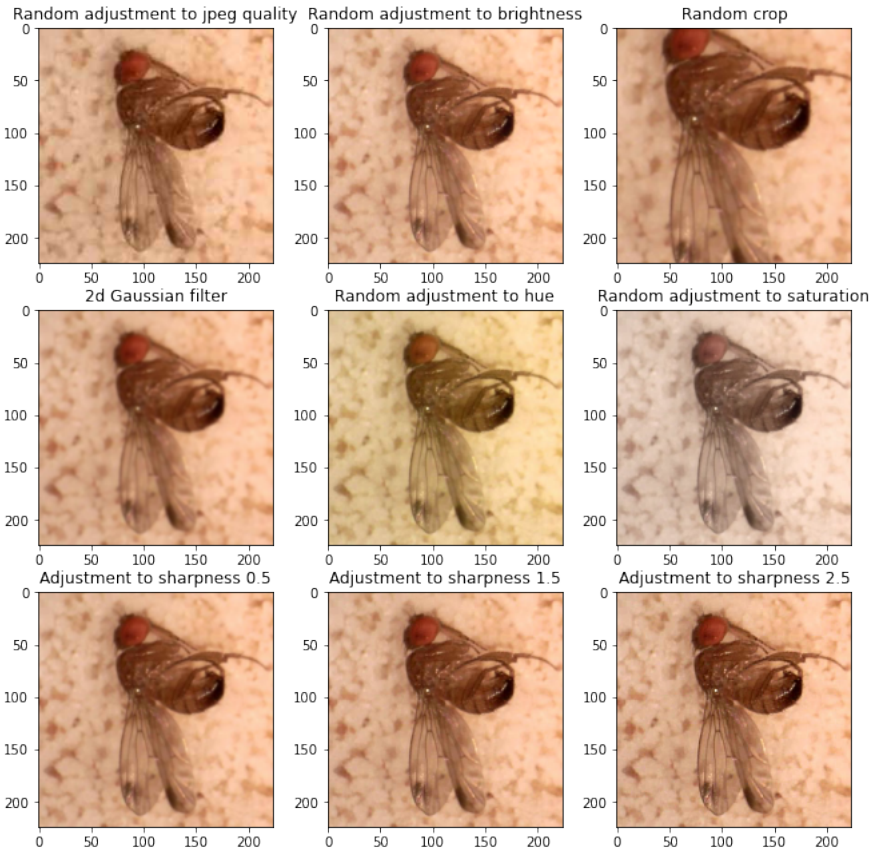

Figure B.2: Shown above are other augmentations which were combined with the background decoupling in the previous figure. The magnitudes of these alterations are exaggerated to show the effects on the example image.

B.2 Pre-clonetool dataset confusion matrix histograms for 32x32 augmentation experiment

One of the implementation problems with an early data set taken for binary classification was the fact that the data used was contaminated by fruit flies of other classes (or sometimes of the same class). This is referred to as the "unlabeled" data set because segmentation data was not created for this set. After training on a higher quality data set the unlabeled data set was still used to examine the entropies of predictions to observe that the majority of predictions had higher entropy. This was done both with and without augmentation.

It is observed that the augmentation was able to reduce the number of false positives and false negatives as well as lower the average entropy of predictions. In other words the model became more confident and more accurate due to augmentation even on the data was highly contaminated with multiple fruit flies on most images in this data set.

This data set was eventually edited with the clonetool from GIMP in order to remove fruit flies which were not intended to be touching the target fruit fly. Using the "cleaner" version of this data edited with clonetool increased the accuracy on validation data substantially.

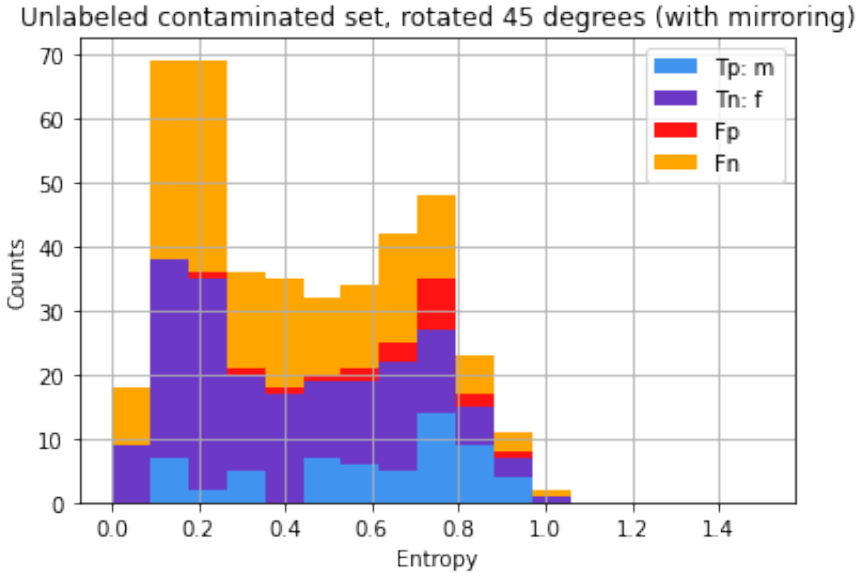

Figure B.3: 32x32 confusion matrix histogram from a model trained on non-augmented data. This test data set had overlapping flies of different classes and was not immediately useful for binary classification. As one might expect the variance in entropy over this distribution is quite large, especially due to the mirroring effect of rotating the whole image. This 45 degree rotation did not use the specialized augmentation but rather the default rotation functions in tensorflow.

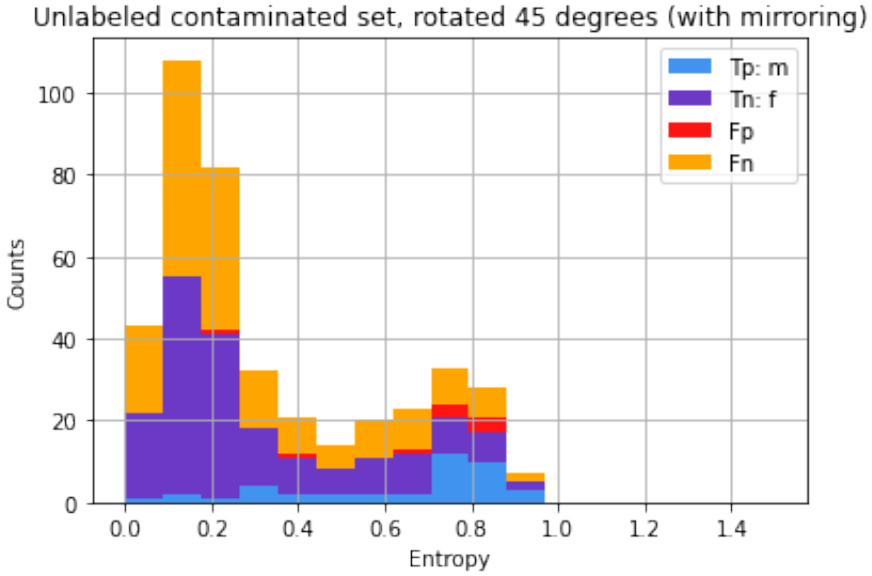

Figure B.4: 32x32 confusion matrix histogram from a model trained on augmented data. The same contaminated data set is shown but this time with the augmentation applied during training. The variance of entropy is noticeably smaller, as the data set is shifted to the left (lower entropy meaning higher confidence in the predictions).

B.3 Difficulty with recall during segmentation

It was observed during experiments that the false negative rate was unexpectedly high. While there were many examples of male fruit flies that had a transparent wing due to the angle of the camera, further inspection revealed that the data set was contaminated with false negatives as well. This was indicated by a steady increase in false negative predictions by the model as the number of epochs increased into over-fitting territory. This was likely part of the reason for the low recall during training of the segmentation data. The segmentation algorithm and its corresponding results was useful in discovering some of the reasons behind the false negative classifications.

In figure B.7 it is most clearly visible that the false negative rate increases steadily. Using validation callbacks during training which collected false negatives and false positives predicted from validation data was useful for troubleshooting the data set. In particular it is notable that using the low entropy predictions smoothed out the learning curve by reducing the tendency of the model to flip flop during training. In many instances for other experiments this low entropy prediction appeared to be more accurate than using raw pixel counts; however this analysis was not exhaustive nor was it within the scope of the project to further investigate the increase in accuracy, as algorithms explicitly designed to make binary predictions such as those in chapter 2 are more suited to evaluate binary accuracy. Nonetheless evaluating the binary accuracy through training of a segmentation algorithm might be useful in determining when to halt training.

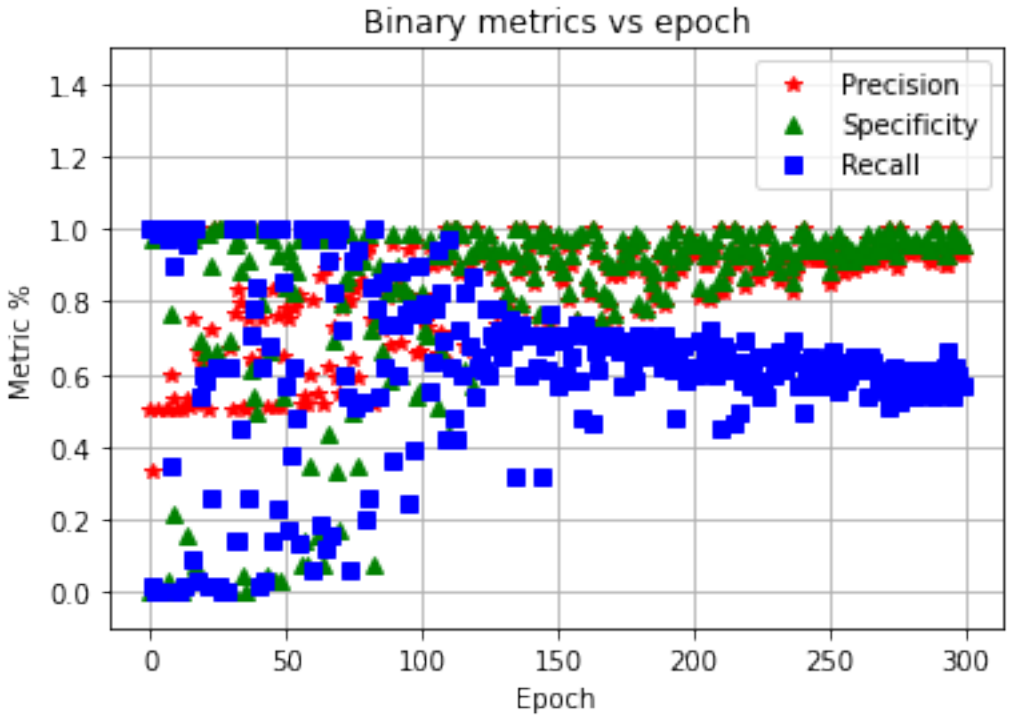

Figure B.5: The precision, specificity and recall are shown while over-fitting out to 300 epochs during the 224x224 experiment. Notice that after about 150 epochs the recall starts to decay while the specificity actually continues slight improvement. This is a strong indication that there are false negatives – male fruit flies contaminating the female fruit fly data set.

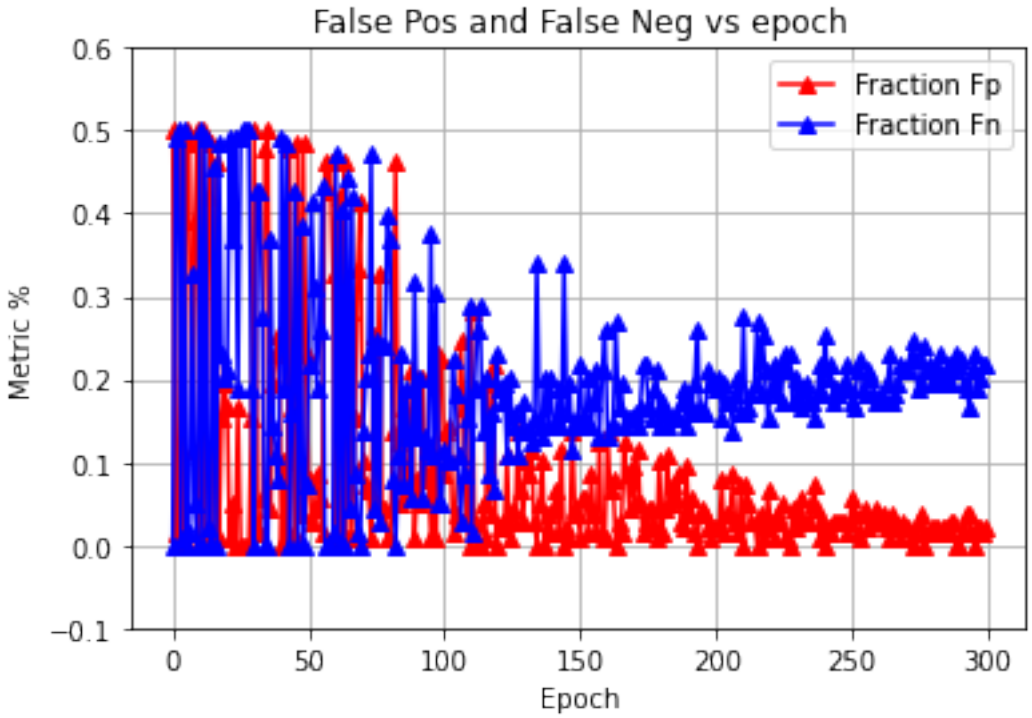

Figure B.6: False negatives and false positives are plotted for the 224x224 overfitting experiment. It can be observed that the model at the start of training predicts predominantly one class, leading to a flip flopping between false positives and false negatives. As the epochs increases beyond approximately 125 epochs the number of false negatives begin to steadily increase.

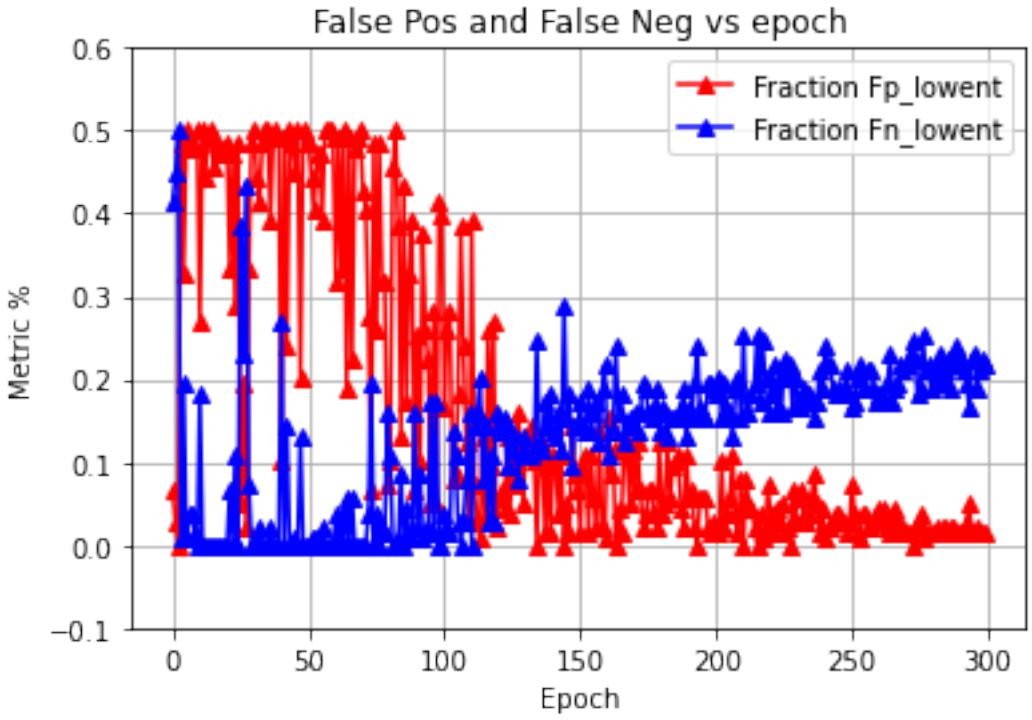

Figure B.7: In the same experiment as the previous figure, low entropy binary classifications are displayed instead of using only the raw pixel count from segmentation as the classification criterion. It appears that the flip flopping is substantially reduced by using low entropy pixels for classification. The steady increase of false negatives is more transparent. This was a key indicator to look for false negatives in the training data, which were later identified and removed.