ALGES: Active Learning with Gradient Embeddings for Semantic Segmentation of Laparoscopic Surgical Images

Josiah Aklilu JOSAKLIL@STANFORD.EDU
Department of Biomedical Data Science
Stanford University
Stanford, California, USA

Serena Yeung SYYEUNG@STANFORD.EDU
Department of Biomedical Data Science
Stanford University
Stanford, California, USA

Abstract

Annotating medical images for the purposes of training computer vision models is an extremely laborious task that takes time and resources away from expert clinicians. Active learning (AL) is a machine learning paradigm that mitigates this problem by deliberately proposing data points that should be labeled in order to maximize model performance. We propose a novel AL algorithm for segmentation, ALGES, that utilizes gradient embeddings to effectively select laparoscopic images to be labeled by some external oracle while reducing annotation effort. Given any unlabeled image, our algorithm treats predicted segmentations as truth and computes gradients with respect to the model parameters of the last layer in a segmentation network. The norms of these per-pixel gradient vectors correspond to the magnitude of the induced change in model parameters and contain rich information about the model's predictive uncertainty. Our algorithm then computes gradients embeddings in two ways, and we employ a center-finding algorithm with these embeddings to procure representative and diverse batches in each round of AL. An advantage of our approach is extensibility to any model architecture and differentiable loss scheme for semantic segmentation. We apply our approach to a public data set of laparoscopic cholecystectomy images and show that it outperforms current AL algorithms in selecting the most informative data points for improving the segmentation model. Our code is available at https://github.com/josaklil-ai/surg-active-learning.

1. Introduction

In recent years, both computer-assisted surgery and the automation of surgical video analysis have shown significant promise in effectively preventing adverse surgical events and postoperative outcomes. The advent of deep neural networks for computer vision has allowed for scalable surgical workflow analysis, skill assessment, and automated surgical feedback for surgical trainees that is unprecedented in surgery (Maier-Hein et al., 2017). Many recent works have adopted convolutional neural networks (CNNs) for recognition tasks on laparoscopic or open surgical video at the frame-level, including tool detection for operative skill assessment, surgical phase recognition, and the detection of anatomical features (Jin et al., 2018; Twinanda et al., 2017; Ban et al., 2021). Automated analysis coming from

© 2022 J. Aklilu & S. Yeung.

deep learning models holds great potential for surgical education, training, and research (Mascagni et al., 2021).

To realize the aforementioned potential of AI in surgery, a detailed visual understanding of the surgical scene over the course of a surgery is necessary. This can be used both to improve the effectiveness of AI-assisted surgery and to ensure surgical safety. A key component of this is the segmentation of anatomical tissue and structures and instrumentation employed during the surgeries. Segmentation is especially relevant for AI in surgery since many works (Madani et al., 2020; Islam et al., 2019; Hasan et al., 2020) utilize the segmentation of the surgical scene as a way to delineate tissue and organ boundaries or clearly identify tools in the surgical scene for either retrospective analysis, real-time corrective feedback, or clinical decision support. For example, Madani et al. (2020) utilize segmentation models to accurately and safely identify dangerous zones of dissection in order to guide clinician decisions during operation.

Therefore, there is active work in developing deep learning based semantic segmentation models (Yang et al., 2017; Gorriz et al., 2017; di Scandalea et al., 2019; Sourati et al., 2018). However, as a consequence of their data hungry nature, collecting a large amount of training data as is standard for training good deep learning models is labor-intensive. The acquisition of large, quality training data sets diverts time and resources away from expert clinicians, which detracts the entire point of utilizing deep learning models for fast, accurate, and scalable computer-assistance and surgical analysis. In general, for most biomedical tasks, acquiring labels for training images requires clinicians with years of background to create the most high quality and reliable image labels (Kim et al., 2020). Carefully labeling data at the frame-level for segmentation of laparoscopic or open surgical videos is not only tedious and expensive, but also unique in that there are highly variant prevalences of different morphological features in the data. Thus, labeling efforts that don't take into consideration the informativeness of data can lead to a data set with redundant samples. For the purposes of developing accurate deep learning models for surgery, it is critical to take into account these considerations as we wish to target larger sets of anatomy and instrumentation for a broader range of surgeries with diverse and complicated workflows.

Active learning (AL) offers the potential of being an effective way of maximizing the performance of deep learning models with minimal available data. Active learning is a framework for machine learning where a model has the ability to beneficially select its own training data (Cohn et al., 1994). In the context of deep learning, a neural network is first trained on a fixed initial training set, and then by some acquisition function, iteratively identifies batches of new data from some unlabeled set that should be labeled by some external oracle (Gal et al., 2017). After labeling, these new data points are included in the training set, and the model is retrained. However, if new data points are chosen without regarding model uncertainty or batch diversity, these new labeled data points may not contribute much to model performance gain (e.g. training a network on 100 diverse images will yield much better model generalizability than training a network on 1000 very similar images). Thus, it is critical in AL to choose a query function that intentionally selects difficult and diverse samples for the model to learn in order to maximize performance.

Several works have applied the active learning paradigm to domains where labeling is costly, such as in biomedical image segmentation (di Scandalea et al., 2019; Sourati et al., 2018; Kim et al., 2020; Yang et al., 2017). However, these approaches require specialized

model architectures effective only on particular data sets that do not generalize well to surgery. In addition to model restrictions, many existing approaches do not consider diversity of acquired samples in each round of AL. Our approach utilizing gradient embeddings encapsulates model uncertainty by measuring how much a new data point influences changes in model parameters. Moreover, these gradient embeddings intrinsically store semantic class information that distinguishes new data points from each other, thus ensuring the heterogeneity of the acquired batch. A few existing approaches have utilized gradient information in a similar way for the classification setting (Ash et al., 2020), but semantic segmentation presents additional challenges that we address such as class imbalance, difficult spatial or structural relationships, and high-dimensionality.

In this work, we propose a novel AL framework called ALGES: Active Learning with Gradient Embeddings for Semantic segmentation. First, we give a new perspective into measuring model uncertainty by observing pixel-level contributions to the gradients of semantic segmentation model parameters with the cross-entropy loss. Specifically, we introduce two ways of summarizing gradients to produce gradient embeddings at the image-level and at the semantic-level. These two approaches confront the problem of high-dimensionality of semantic data, a challenge that does not exist in the classification setting. We note that the L2-norms of these gradient embeddings can be used as a measure of uncertainty for data points that need to be selected for labeling. ALGES relies on the core idea that these gradient embedding magnitudes encode uncertainty information in a robust manner. Furthermore, we compare how these methods of embedding generation perform when used as seeds to the k-MEANS++ initialization strategy (Arthur and Vassilvitskii, 2007), which ensures diversity of the acquired batch of uncertain data points. Finally, we apply ALGES to a publicly available data set of laparoscopic surgery images (Hong et al., 2020), where class imbalance and semantic difficulty are key issues. In summary, our contributions are:

- We develop a new AL framework that leverages gradient information at the semantic level in a way which to our knowledge, has not been previously explored.
- We introduce two methods of deriving gradient embeddings for segmentation at the image-level and semantic-level for unlabeled images that can be used to acquire diverse batches of data with high model uncertainty.
- ALGES displays performance increases over established AL algorithms on a segmentation data set of laparoscopic cholecystectomy images (Hong et al., 2020) that is a sample of real-world surgical video data.
- Our proposed method is generalizable to model architecture and data setting, making ALGES a prime candidate for querying data to be annotated in not only laparoscopic surgery, but a broad variety of biomedical domains, where unlabeled data is abundant and highly heterogeneous. ALGES is adaptable, simple, and effective.

Generalizable Insights about Machine Learning in the Context of Healthcare

This work highlights the significance of understanding *how* model parameters change so that the best training data set can be assembled under a limited budget. Deep learning models are trained with gradients, and data points that induce large gradients are ones

which the model has to make large updates to its parameters in order to learn. Thus, gradients intrinsically hold uncertainty information and should be incorporated in approaches designed to improve model performance with limited data. ALGES is one of the first approaches to utilizing gradients of segmentation losses in the active learning setting, and this idea could be useful in creating frameworks for other computer vision tasks on biomedical data. This work outlines a fruitful direction for research in minimizing labeling efforts to spare valuable clinician time and resources, while also improving model performance for AI-assisted surgery in the enhancement of surgical healthcare.

2. Related Work

Many works have investigated the active learning paradigm in the context of computer vision. However, the breadth of literature is relatively limited for active learning for segmentation as opposed to image classification. Here, we review the current state of active learning research and the application of this framework in biomedical image segmentation.

There are several established AL approaches in the classification setting, such as Max Entropy sampling, Margins sampling, and Least confidence sampling, which utilize a model's prediction probability as a measure of uncertainty for unlabeled data points, of which a batch of highly uncertain samples are acquired (Shannon, 1948). The Bayesian active learning by disagreement algorithm incorporates ideas from Bayesian modeling in deep learning in approaching the AL framework (Gal et al., 2017). Newer approaches to AL for classification problems have emerged where the diversity is considered in order to reduce redundant data points in acquired batches (Kirsch et al., 2019; Hemmer et al., 2020; Sener and Savarese, 2018). In this manuscript, we focus on the more challenging semantic segmentation setting.

There has been growing work in developing AL frameworks for the nuanced semantic segmentation space. Yang et al. (2017) proposed a new fully convolutional neural network architecture that performs state-of-the-art in gland and lymph node segmentation, but their AL algorithm hinges on this highly specialized model architecture and is not robust to other architectures that perform better on other biomedical data sets. Sourati et al. (2018) used Fisher information as a criterion for querying the best images for patch-wise segmentation of both adolescent and newborn brains. Their work demonstrates how gradients of model parameters are used to infer the value of an image being labeled, however they aggregate gradients in a limiting way to reduce the parameter space when computing the Fisher information, which hinders the ability to capture rich spatial information in the multi-class segmentation setting. di Scandalea et al. (2019) similarly tries to leverage uncertainty as a criterion by using dropout regularization in an axon-myelin segmentation model, while Mackowiak et al. (2018) take a unique approach incorporating annotation cost when creating measures for selecting regions for annotation, although both methods do not address diversity when acquiring regions for annotation, which could lead to redundant data.

A few more recent works try to capture the notion of semantic difficulty in their approaches. Siddiqui et al. (2020) teaches a novel framework that utilizes RGB-D data to simulate entropy and divergence of multiple vantage points in order to query regions for annotation. This approach is impractical in the surgery setting and many other biomedical domains, since pose and depth information is not readily available with this type of data. Xie et al. (2020) also introduce a unique probability attention module that ensures that

a similar semantic difficulty value is computed for pixels of a particular semantic. This semantic difficulty is guided by an error mask and is combined with uncertainty information to rank images that should have the highest priority for annotation. We compare our method with this work in subsequent sections.

Other works attempt to mitigates various problems arising in larger batch settings in active learning. Ghorbani et al. (2021) approach scalability in AL by filtering data points with the highest data Shapley values in the unlabeled pool before applying traditional AL methods in an attempt to reduce the unlabeled data space. Citovsky et al. (2021) propose an efficient active learning algorithm for dealing with the batch sizes several orders of magnitude larger than typical settings, although we focus on the small batch setting as is common for AL for small biomedical data sets.

A key preliminary work (Ash et al., 2020) to this manuscript proposes an algorithm for batch active learning in the classification setting using gradient embeddings. These embeddings are computed as a scaling of normalized activations from the penultimate layer of any neural network. They show that the magnitude of the gradients of the parameters in the last layer of a network implicitly contains information on a data point's influence on the model, and thus can be used as a natural measure of model uncertainty. However, computing and utilizing gradient embeddings in this manner is not readily apparent in the segmentation setting. Our work introduces two methods, one considering gradients for the image as a whole, and one considering gradient contributions from each pixel along with semantics, to show how gradient embeddings can be effectively used in active learning for semantic segmentation.

3. Methods

3.1. Preliminaries

We provide an overview of ALGES in Figure 1. In the subsequent sections, we clearly define the AL problem and show two methods of deriving gradient embeddings for segmentation.

3.1.1. Active learning

For completeness, we explicitly formulate active learning in the batch setting. Let $\mathscr{D} = \mathcal{D}_L \cup \mathcal{D}_U$ denote the entire set of data points, $\mathcal{D}_L$ denote labeled data, and $\mathcal{D}_U$ unlabeled data. Given $\mathscr{D}$, a model $\mathcal{M}$ trained on $\mathcal{D}_L$, and a batch size B, the AL objective for each round is to select B images x^* from $\mathcal{D}_U$ using a query function q:

$$\{x_1^*, \ldots, x_B^*\} = q(\{x : x \in \mathcal{D}_U\}, \mathcal{M})$$

where q can be any function representing a criterion for selection. Concretely, the AL framework selects a batch of data points from the unlabeled data pool via some criterion utilizing the current trained model. A simple AL objective function is the random sampling method, which is equivalent to selecting B data points uniformly at random from the unlabeled pool. There are other common acquisition functions used to select batches of unlabeled data for labeling, and we compare our method to these standards in subsequent sections.

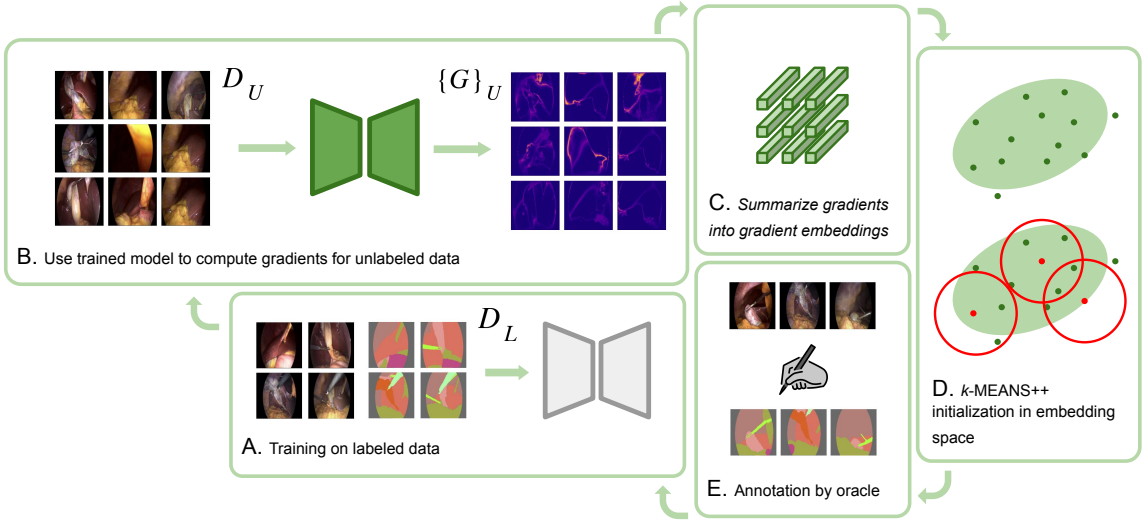

Figure 1: Workflow of ALGES. A) First, the segmentation model is trained on the limited initial pool of training data which have labels, $\mathcal{D}_L$. B) The trained model then makes predictions on the unlabeled data $\mathcal{D}_U$ and computes gradients $\mathbf{G}$ for each image as outlined in this manuscript. C) The gradients $\mathbf{G}$ per image are summarized in the two methods we demonstrate in this work into gradient embeddings g per image. D) The centroids of these embeddings are found using k-MEANS++ initialization, resulting in a diverse and representative batch of images. E) The selected images are annotated by a clinical expert or other external oracle and added to the training pool of labeled data $\mathcal{D}_L$, where the model is retrained and the entire process is repeated until the model achieves the desired performance.

3.1.2. Segmentation model, inputs, and outputs

Let lowercase bold letters denote vectors and uppercase bold letters denote matrices. Let f represent a neural network with parameters $\theta = (\mathbf{W}, \mathbf{V})$ where $\mathbf{W}$ denotes the parameters of the last layer of the network and $\mathbf{V}$ denotes the parameters of all other layers in the network. Then $f(\mathbf{x}, \theta) = \sigma(\mathbf{W} \cdot z(\mathbf{x}, \mathbf{V}))$, where σ denotes a nonlinearity (i.e. softmax, $\sigma(x)_i = e^{x_i} / \sum_{k=1}^{K} e^{x_k}$) and $z(x, \mathbf{V})$ represents activations coming from the penultimate layer of the network.

In this work, we consider a segmentation model where the weights $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_K)^\top \in \mathbb{R}^{K \times d}$ of the last layer of the network are the parameters of a 1×1 convolution layer with K filters, where K is the number of output classes and each filter $\mathbf{w}_k \in \mathbb{R}^{d \times 1}$. We denote a singular input instance to this model as $\mathbf{X} \in \mathbb{R}^{3 \times N \times M}$, where N and M are the dimensions of the image, and we signify the ground truth output as $\mathbf{Y} \in \mathbb{R}^{K \times N \times M}$ with the same height and width dimensions as the input. Let $\mathbf{x}_{ij} \in \mathbb{R}^3$ indicate a singular pixel in the input, where $1 \leq i \leq N$ and $1 \leq j \leq M$, and $\mathbf{y}_{ij} \in \mathbb{R}^K$ a one-hot encoding of the corresponding ground truth classification of this pixel. Finally, we can define the predictions of

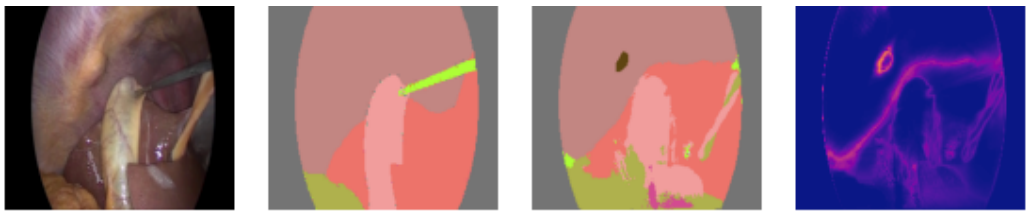

Figure 2: (a) The input, in our application a laparoscopic image. (b) Ground-truth segmentation of the input delineating relevant anatomy and instruments. (c) Current model predictions. (d) Gradients induced by this output. The magnitudes in the heat map are L2-norms of the gradient embeddings per pixel.

this model as $\hat{\mathbf{Y}} \in \mathbb{R}^{N \times M}$ where each pixel has prediction $\hat{y}_{ij} = \text{argmax}_{k \in \{1,\ldots,K\}} p_{ijk}$, and where $\mathbf{p}_{ij} = f(\mathbf{x}_{ij}, \theta) \in \mathbb{R}^K$ are the predicted class probabilities for an individual pixel.

We note that semantic segmentation of any image can be thought of as multiple multi-class classification problems, where the input for each problem is one pixel with 3 RGB values and the output is some one-hot encoding representing the predicted semantic class of that pixel.

3.1.3. Computing Gradients for Segmentation with Cross Entropy Loss

Ash et al. (2020) utilize gradient embeddings in the classification setting to acquire batches of unlabeled data. They compute gradients a for the final fully connected layer in a neural network. We observe that segmentation models typically have dense output, and many have 1×1 convolutions convolved over penultimate activations. This is equivalent to multiple smaller classification problems where the convolution filters are like the weights of a fully connected layer. We derive gradient expressions with respect to segmentation and will utilize these when we present our approach in subsequent sections. We use these calculated gradients to produce embeddings at the image-level and then explore a different approach that incorporates semantic information not readily present at the pixel level.

Consider the semantic segmentation setting. A commonly used loss function in optimizing segmentation networks is the standard cross-entropy loss. Since we do not have labels in the active learning setting at first, we must treat the current model's predictions of the output as ground truth. The cross-entropy loss for one image is then:

$$\begin{aligned} \mathscr{L}_{CE}(f(\mathbf{X}, \theta), \hat{\mathbf{Y}}) &= \frac{1}{NM} \sum_{ij} \ell_{ij} \\ &= \frac{1}{NM} \sum_{ij} \left(-\sum_{k=1}^{K} \hat{y}_{ijk} \ln p_{ijk} \right) \\ &= \frac{1}{NM} \sum_{ij} \left(\ln \left(\sum_{k=1}^{K} e^{\mathbf{w}_k \cdot \mathbf{z}_{ij}} \right) - \mathbf{w}_y \cdot \mathbf{z}_{ij} \right) \end{aligned} \tag{1}$$

which can be thought of as an average of per-pixel losses. We can then use the expression for the gradients corresponding to any output label k computed in Ash et al. (2020) for

an individual pixel, since the classification of a pixel is a multi-class classification problem. Substituting into the segmentation formulation above:

$$\begin{aligned}\frac{\partial \mathscr{L}_{CE}}{\partial \mathbf{w}_k} &= \frac{1}{NM}\sum_{ij}\frac{\partial}{\partial \mathbf{w}_k}\left(\ln\left(\sum_{k=1}^{K} e^{\mathbf{w}_k\cdot\mathbf{z}_{ij}}\right) - \mathbf{w}_y\cdot\mathbf{z}_{ij}\right) \\ &= \frac{1}{NM}\sum_{ij}(p_{ijk} - \mathbb{I}(k=\hat{y}_{ij}))\mathbf{z}_{ij} = \frac{1}{NM}\sum_{ij}\left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_k}\right)_{ij} \in \mathbb{R}^d \\ \frac{\partial \mathscr{L}_{CE}}{\partial \mathbf{W}} &= \left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_1},\ldots,\frac{\partial \mathscr{L}}{\partial \mathbf{w}_K}\right)^\top\end{aligned} \tag{2}$$

where $\left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_k}\right)_{ij} = (p_{ijk} - \mathbb{I}(k=\hat{y}_{ij}))\mathbf{z}_{ij}$ is an individual pixel's contribution to the gradient calculation for weights $\mathbf{w}_k$. Note that each $\mathbf{w}_k$ is d dimensional, and they represent weights for a particular output class. There are $(N \times M) \times K$ such terms, which we can interpret as a matrix:

$$\mathbf{G} = \begin{bmatrix} \left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_1}\right)_{11} & \cdots & \left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_k}\right)_{11} & \cdots & \left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_K}\right)_{11} \\ \vdots & & \vdots & & \vdots \\ \left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_1}\right)_{ij} & \cdots & \left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_k}\right)_{ij} & \cdots & \left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_K}\right)_{ij} \\ \vdots & & \vdots & & \vdots \\ \left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_1}\right)_{NM} & \cdots & \left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_k}\right)_{NM} & \cdots & \left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_K}\right)_{NM} \end{bmatrix} \in \mathbb{R}^{(N\cdot M)\times K\times d} \tag{3}$$

This matrix $\mathbf{G}$ represents each pixel $\mathbf{x}_{ij}$'s contribution to the gradients of the parameters in the last layer of the model.

3.2. Our approach

Our approach to active learning for multi-class semantic segmentation is detailed in Algorithm 1. We first randomly sample B images from the available unlabeled data $\mathcal{D}_U$ and train a model on this initial batch of images. Then, for each iteration of AL, we employ the current model to make predictions on all remaining unlabeled data and compute gradient embeddings in two different approaches, at the image-level and at the semantic-level, which we describe in the next sections. The resulting embeddings are used as a seeding to the k-MEANS++ initialization algorithm, which selects centroids of these unlabeled data embeddings that are consistently of high magnitude and high diversity (Arthur and Vassilvitskii, 2007). The selected unlabeled data points are then sent to an external oracle for labeling and then consumed in the labeled data set, and the process is repeated again with a model trained on this slightly larger training data set.

3.2.1. Image-level Gradient Embeddings

We investigate computing gradient embeddings per-image in two ways. The first directly computes induced change in the weights $\mathbf{W}$ due to the entire image, which is a more direct extension of Ash et al. (2020). To compute gradient embeddings $g_{\mathbf{X}}$ per image, we consider the last convolutional layer weights $\mathbf{W}$ in the segmentation model as the weights of a fully

Algorithm 1 ALGES: Active Learning with Gradient Embeddings for Segmentation

Require: A segmentation network $f(\mathbf{X}, \theta)$, an unlabeled data pool $\mathcal{D}_U$, number of active learning rounds R, and batch size B.

Randomly sample some initial pool $\mathcal{D}_L$ of training data from $\mathcal{D}_U$ with labels.
Train model f with parameters θ_1 on $\mathcal{D}_L$.
for $r \leftarrow 1$ **to** R **do**
    for all $\boldsymbol{X} \in \mathcal{D}_U$ **do**
        1. Get prediction on $\mathbf{X}$, $\hat{\mathbf{Y}}$.
        2. Compute the gradient embedding $g_{\mathbf{X}}$ by Eq. 4 or Eq. 6.
    end
    Use these embeddings in the k-MEANS++ initialization algorithm with B centroids to define a subset $\mathcal{D}_B \in \mathcal{D}_U$ of B images.
    $\mathcal{D}_L \leftarrow \mathcal{D}_L \cup \mathcal{D}_B$
    $\mathcal{D}_U \leftarrow \mathcal{D}_U \backslash \mathcal{D}_B$
    Train the model on $\mathcal{D}_L$ to get updated parameters θ_{r+1}.
end
return θ_{R+1}

connected layer. We can then define the k-th block of the gradient embedding for one image as:

$$(g_{\mathbf{X}})_k := \frac{\partial \mathscr{L}}{\partial \mathbf{w}_k}^\top = \frac{1}{NM} \sum_{ij} \mathbf{G}_k \in \mathbb{R}^d \tag{4}$$

which is an average of the vector values across columns of $\mathbf{G}$. Intuitively, the resulting $K \times d$ dimensional vector summarizes gradient information from every pixel for each class. The L2-norm each element of $(g_{\mathbf{X}})$ is a measure of how much an input image $\mathbf{X}$ influences the model's parameters for a specific class. This is an adjunct formulation of the embeddings described in Ash et al. (2020).

3.2.2. Semantic-level Gradient Embeddings

We want to consider the contribution of pixels to the gradients to incorporate local semantic information in these embeddings, so we propose a different approach in using the gradients $\mathbf{G}$. We stress that a 1×1 convolution over pixels is fundamentally multiple multi-class classification problems. This means we can view *rows* of $\mathbf{G}$ as gradient embeddings for each pixel. Precisely, for one pixel $\mathbf{x}_i j$:

$$\begin{aligned}(g_{\mathbf{x}_{ij}}) = \mathbf{G}_{ij} &= \left(\left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_1} \right)_{ij}, \cdots, \left(\frac{\partial \mathscr{L}}{\partial \mathbf{w}_K} \right)_{ij} \right)^\top \\ &= ((p_{ij1} - \mathbb{I}(1 = \hat{y}_{ij}))\mathbf{z}_{ij}, \cdots, (p_{ijK} - \mathbb{I}(K = \hat{y}_{ij}))\mathbf{z}_{ij}) \in \mathbb{R}^{K \times d}\end{aligned} \tag{5}$$

The Frobenius norm of these embeddings conservatively estimate the pixel's point influence on the current model. Consequently, these norms are good estimators of model uncertainty. We can then imagine a heat map as in Figure 2. Aggregating norms for pixels

of the same semantic class predicted by the current model yields elements of a semantic embedding.

$$(g_{\mathbf{X}})_k := \sum_{\substack{ij \\ \hat{y}_{ij}=k}} \mathbf{G}_{ijk} \in \mathbb{R}^d \tag{6}$$

This is equivalent to calculating the Frobenius norm of the embedding for an individual pixel (Eq. 6), and summing these norms for pixels of the same predicted class (e.g. gallbladder). The elements of the resulting K dimensional vector $(g_{\mathbf{X}})$ contain rich semantic difficulty information per semantic class.

3.2.3. Selection by k-MEANS++ initialization

Once we've compute gradient embeddings for each image in the unlabeled pool using one of the aforementioned methods, we wish to select a batch of data points so as to maximize diversity of the batch, while also maximizing the total uncertainty of that batch. A way of choosing data points (centers) that would approximately cluster the data set in the most representative manner is the k-MEANS++ initialization scheme (Arthur and Vassilvitskii, 2007). A seeding for the popular k-MEANS clustering algorithm, the k-MEANS++ initialization, tends to select data embeddings that are diverse and of high-magnitude, as noted in Ash et al. (2020). We use this initialization as a way of selecting diverse and high-magnitude gradient embeddings, which correspond to diverse images that the model has low predictive confidence.

4. Experiments on Laparoscopic Surgical Data

In this work, we demonstrate the effectiveness of ALGES on a real-world, publicly available laparoscopic surgical data set called *CholecSeg8k* (Hong et al., 2020). *CholecSeg8k* was developed for training better computer vision models in the laparoscopic surgery space for various downstream tasks, including computer-assisted surgery, operative skill assessment, and the enhancement of surgical safety. It is derived from the *Cholec80* data set, a collection of 80 videos of full laparoscopic cholecystectomy surgeries performed by 13 different surgeons recorded at 25 fps. 8080 image frames were extracted from 17 videos in *Cholec80* to form *CholecSeg8k*. There are 13 semantic classes that delineate relevant instruments and anatomy in cholecystectomies: abdominal wall, liver, gastrointestinal tract, fat, grasper, connective tissue, blood, cystic duct, L-hook electrocautery, gallbladder, hepatic vein, liver ligament, and a black background class. Each image frame is 854×480 pixels.

As noted by Hong et al. (2020), there is a severe class imbalance of semantic areas in the images present in *CholecSeg8k*, making this a difficult data set to segment, but representative of real-world surgical video data. We note that since the frames are extracted from only 17 videos, many of the images in the data set are highly similar due to their shared origin. To ensure that our trained models do not see test data similar to the data they were trained on, we enforce constraints in our train, validation, and test splits. We divide *CholecSeg8k* into 4640 frames for training, 1600 frames for use in validation, and 1640 frames for final testing. This split was chosen with the constraint that frames from a single video could

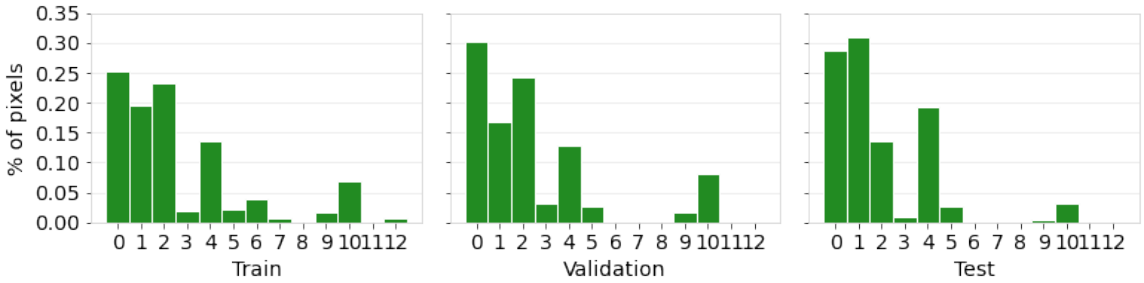

Figure 3: Label distribution of data split for *CholecSeg8k*. Severe class imbalance and difference in distribution implicate semantic segmentation difficulty, but reflect real-world surgical video data.

only be present in one split. This amounts to having 8 videos in the training split, 4 videos in validation, and 5 videos in the test split. We perform a simple resizing transformation to the data so that all frames have square dimensions 224×224. Figure 2 illustrates the pixel label distribution of each split after the aforementioned data processing.

4.1. Main Results

In our experiments, we use a U-Net (Ronneberger et al., 2015) with a ResNet-18 backbone (He et al., 2016), where the last convolution was modified for 13 output classes. For every round of active learning, we train the model for 50 epochs with a learning rate of $5e^{-}5$ and using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ parameters. We use an initial batch of 10 images, and then query 10 images per round of AL to observe differences in AL algorithms in the small batch, low data setting. We perform 30 iterations of AL, and we use the mean of 3 experiments with different random seeds to report mean test Dice score (Milletari et al., 2016). We compare our algorithm with several other AL algorithms in the literature modified for semantic segmentation (see Figure 3):

- *Random*: The baseline AL algorithm which randomly acquires a select number of images from the unlabeled data pool for annotation.
- *Uncertainty based methods*: AL algorithms that select unlabeled data for annotation based on some measure of model uncertainty (see Appendix B). *Max Entropy*: A method that computes the predictive entropy of a sample as a measure of model uncertainty and selects the unlabeled data points with the highest value. *Margin sampling*: An approach that computes the difference in the two largest predicted class probabilities for a data point and that selects points with the smallest margins. *Least confidence*: Selects points that have the smallest max predicted class probability.
- *Core-set*: A diversity based approach aimed at the batch setting of active learning. We implement the greedy approach outlined in Sener and Savarese (2018) that approximates the k-Center problem by iteratively selecting new data points such that the largest distance from them to the nearest centers in the existing data are minimized.

The method has been shown to be effective for active learning for convolutional neural networks in the batch setting.

- *DEAL*: A recent approach proposing an AL method for segmentation based on a probability attention module that generates a semantic difficulty map for an image (Xie et al., 2020). We compare against the difficulty entropy version of their algorithm.
- *ALGES-img*, *ALGES-sem*: Our algorithm using Eq. 4 or Eq. 6 to compute gradient embeddings, respectively.

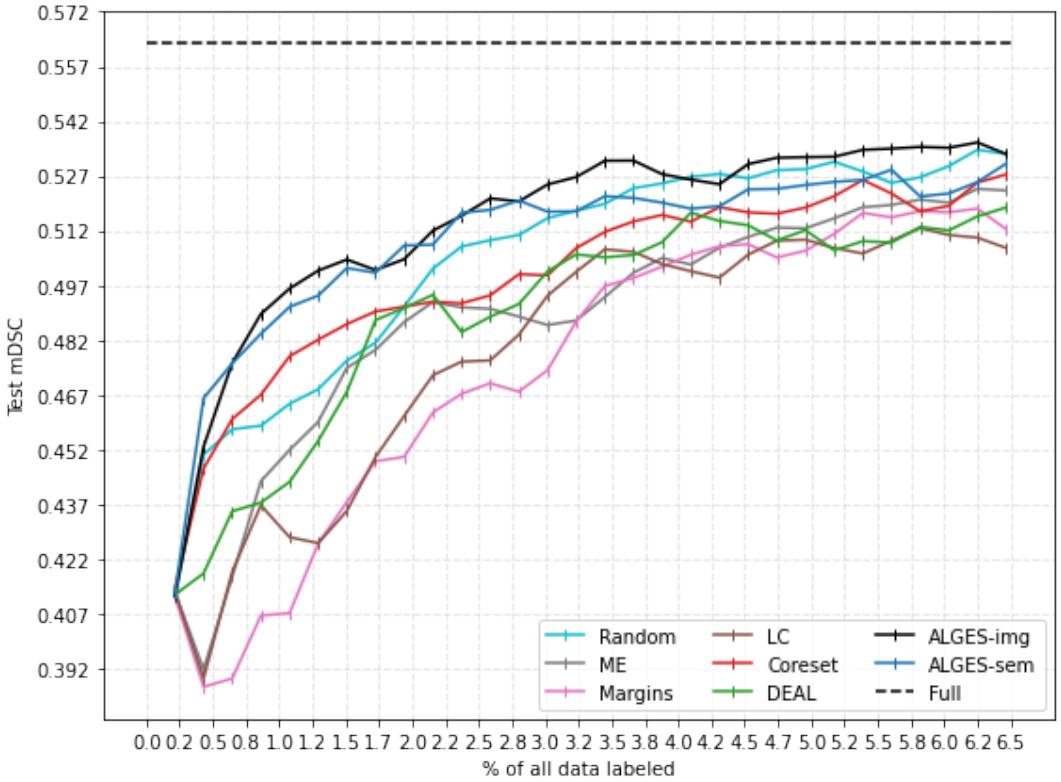

Figure 4: *Left*: Active learning curves for different algorithms. ME stands for Max Entropy, LC is Least confidence sampling. Full refers to training the model under full supervision with the entire available training data. Our method using image-level gradient embeddings (ALGES-img) outperforms other algorithms in selecting informative images for annotation by a clinical expert.

Our methods using gradient embeddings derived from both Eq. 4 and Eq. 6 outperform other strategies in nearly all iterations of AL up to 6.5% of our total data pool being labeled. Methods relying on uncertainty alone (i.e. Max Entropy, Margin sampling, Least

Table 1: Segmentation performance of U-Net model with various percentages of all data labeled (trained with). The relevant metric is mean Dice score. ALGES-img performs better than all methods in small batch, low data settings.

Method	0.43% labeled	1.94% labeled	3.02% labeled	4.96% labeled
Random	0.448	0.480	0.516	0.522
Coreset	0.431	0.484	0.496	0.519
DEAL	0.389	0.476	0.496	0.517
ALGES-img	**0.458**	**0.522**	**0.531**	0.532
ALGES-sem	0.419	0.498	0.526	**0.537**

confidence) perform poorly especially in low data regions (less than 5% of total images labeled). We hypothesize this is due to the fact that these methods neglect optimizing diversity in the batch setting, and so they select redundant data points that although the model has difficulty making predictions with, are highly similar. This can even lead to performance decreases as the model attempts to fit redundant data. The recently proposed DEAL algorithm performs similarly to random acquisition, indicating the ineffectiveness of the probability attention module in capturing segmentation difficulty for surgical data. Coreset also doesn't perform up to par with our method even though it considers data diversity and at its core tries to solve a similar maximum data cover by centers problem. This may be due to the way embeddings are generated for the Coreset algorithm, which lack semantic information. We tabulate specific Dice score values in Table 1.

We also note the difference between using the image-level and semantic-level gradient embeddings. Figure 4 also demonstrates the relative performance of these methods. Although AL using image-level embeddings consistently outperforms AL using the semantic-level embeddings and in later rounds of AL, there is not a significant difference between the two methods in earlier rounds on average (low data region).

5. Discussion

Interpreting gradient embeddings. Our primary inspiration for the use of gradient embeddings is how they encapsulate uncertainty information while providing a way to ensure diversity in acquired batches during AL. Deep learning models are trained via gradient updates, and data points that bring about large gradient updates are data points that the model should learn. Thus, uncertainty for an image can be captured by computing gradients of the weights for segmentation models, and representing these gradients as an image-level embedding. We have also shown that at the semantic level, considering pixels of a particular semantic class that induce large updates to the model parameters gives us insight into model uncertainty for particular regions of an image. We substantiate this intuition with empirical studies, but we recognize that this frame of reference can provide meaningful insight in understanding uncertainty for deep learning models.

The AL approaches that completely depend on model uncertainty as a criterion for selecting images to be labeled are poor AL algorithms since multiple semantically similar images can have the same predictive uncertainty from a model. Diversity based approaches better ensure the value of acquired data in improving model performance. However, additionally summarizing semantic information in embeddings in a deliberate manner like with our semantic-level gradient embeddings ensures morphological diversity in acquired batches that is critical for developing segmentation models for biomedical datasets.

The uniqueness of the semantic segmentation problem adds another dimension to this investigation. Although pixel-wise segmentation is a group of classification problems, relationships between these classifications at the semantic level are the key components of producing effective algorithms with superior segmentation quality. The use of gradient embeddings in the manner outlined in this work incorporates these semantic relationships in a resourceful way, since the gradients are derived from convolutional filters that inherently aggregate information from receptive fields of penultimate layer activations. In addition, by summarizing gradients at the semantic-level into a gradient embedding, we also provide an avenue for ensuring diversity between acquired data points.

Clinical value. In this work, we explore our AL framework in the context of deep segmentation models for laparoscopic surgery. High quality segmentation of the surgical scene can offer significant utility for AI-assisted surgery. For example, segmentation of instruments can be used to track tools and prevent certain tool movements that could cause adverse events (i.e. aiding surgeons by reducing physiologic tremor) in robotic assisted minimally invasive surgery (Azqueta-Gavaldon et al., 2020). Segmentation of tissue and organ boundaries can also be used to identify dangerous zones of dissection in order to guide clinician decisions during operation (Madani et al., 2020). Recent studies such as one by Hasan et al. (2020) particularly emphasize how segmentation is imperative for intra-operative clinical decision support. Intra-operative segmentation of tools, for example, can be used as a preprocessing step for inpainting instrumentation through the view of the laparoscope (i.e. making instruments appear transparent through the camera view). This allows clinicians to easily see tissue structures obscured by instruments for better execution of surgical procedures. However, deep segmentation models need to be trained on biomedical data sets particular to the clinical setting in which they are deployed.

Biomedical datasets in general are highly diversified and usually contain less samples than other datasets used for training deep learning models. Laparoscopic surgical datasets in particular, like those developed for studies by Madani et al. (2020); Maqbool et al. (2020), are difficult to construct, necessitating annotators with years of clinical expertise. Images extracted from laparoscopic videos from differing institutions have different lighting, instrumentation, modes of video capture, and various differences in quality. Non-laparoscopic surgeries are even more varied, underlining the need for larger datasets from a wide variety of sources to develop more generalizable models that can capture pathology or anatomy that are not seen during train time (Madani et al., 2020).

AL frameworks significantly reduce annotation efforts in assembling biomedical datasets for deep learning models. The best AL algorithms select a diverse batch of data points that are most informative to the current model, ensuring the quality of data being labeled by annotators. This is paramount when attempting to build larger biomedical datasets so that annotation resources are efficiently used to obtain maximum data value. Table 1 highlights

how our method results in the best model performance when only a small percentage of available unlabeled data is labeled. In addition, our method is effective in the low batch setting, making our approach suitable for situations where data is limited and there is a low annotation budget or high annotation cost. This is the primary setting of biomedical computer vision datasets not only in surgery, but in pathology, radiology, and other domains.

It would be remiss to ignore limitations of our approach. Although not explored in the scope of this work, an AL framework that allows for selecting areas of images for annotation as opposed to entire images could further reduce annotation efforts for training accurate computer vision models. We hypothesize that only selecting regions of an image where gradient embedding magnitudes are large would still an effective AL strategy and may even enable the use of more unlabeled data, albeit at the expense of computational cost. Additionally, we could more explicitly target model performance on rarer segmentation classes, which could help mitigate the imbalance problem present in many biomedical datasets, especially in those domains where abnormalities are pervasive.

6. Conclusions

In this manuscript, we propose a novel active learning algorithm for semantic segmentation using gradient embeddings (ALGES). We present a derivation of pixel-level contributions to gradients for the segmentation setting, treating segmentation as multiple multi-class sub-problems. We also provide insight into gradient norms and their function as measures of model uncertainty, and we introduce two approaches to building gradient embeddings. We demonstrate the effectiveness of our algorithm compared to other AL algorithms on a dataset of laparoscopic cholecystectomy images and highlight its efficacy in low data regimes, making it a prime candidate for domains where annotation cost is high. For future work, we plan to explore AL frameworks that allow for region selection that utilize gradient embeddings in an effective way. Our works suggests that investigating intrinsic elements of deep learning models such as gradient embeddings is a fruitful direction for research in developing algorithms for maximizing data utility.

Acknowledgments

This work was partially supported by an Intermountain-Stanford Collaboration Grant, the National Science Foundation under Grant No. 2026498, and the Stanford Clinical Excellence Research Center. J.A. was also funded by a Stanford Graduate Fellowship.

References

David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA, 2007. Society for Industrial and Applied Mathematics.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

Iñigo Azqueta-Gavaldon, Florian Fröhlich, Klaus H. Strobl, and Rudolph Triebel. Segmentation of surgical instruments for minimally-invasive robot-assisted procedures using generative deep neural networks. *CoRR*, abs/2006.03486, 2020.

Yutong Ban, Guy Rosman, Thomas M. Ward, Daniel A. Hashimoto, Taisei Kondo, Hidekazu Iwaki, Ozanan R. Meireles, and Daniela Rus. Aggregating long-term context for learning laparoscopic and robot-assisted surgical workflows. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14531–14538, 2021.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

David Cohn, Zoubin Ghahramani, and Michael Jordan. Active learning with statistical models. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994.

Melanie Lubrano di Scandalea, Christian S. Perone, Mathieu Boudreau, and Julien Cohen-Adad. Deep active learning for axon-myelin segmentation on histology data, 2019.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1183–1192. JMLR.org, 2017.

Amirata Ghorbani, James Y. Zou, and Andre Esteva. Data shapley valuation for efficient batch active learning. *ArXiv*, abs/2104.08312, 2021.

Marc Gorriz, Axel Carlier, Emmanuel Faure, and Xavier Giró-i-Nieto. Cost-effective active learning for melanoma segmentation. *CoRR*, abs/1711.09168, 2017.

S. M. Kamrul Hasan, Richard Simon, and Cristian Linte. Segmentation and removal of surgical instruments for background scene visualization from endoscopic / laparoscopic video. volume 11598, 11 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Patrick Hemmer, Niklas Kühl, and Jakob Schöffer. Deal: Deep evidential active learning for image classification. *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 865–870, 2020.

W.-Y. Hong, Chang-Lung Kao, Y.-H. Kuo, J.-R. Wang, W.-L. Chang, and C.-S. Shih. Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *ArXiv*, abs/2012.12453, 2020.

Mobarakol Islam, Daniel Atputharuban, and Ravikiran Ramesh. Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. *IEEE Robotics and Automation Letters*, PP:1–1, 02 2019.

Amy Jin, Serena Yeung, Jeffrey Jopling, Jonathan Krause, Dan Azagury, Arnold Milstein, and Li Fei-Fei. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 691–699. IEEE, 2018.

Taehun Kim, Kyung Hwa Lee, Sungwon Ham, Beomhee Park, Sangwook Lee, Dayeong Hong, Guk Bae Kim, Yoon Soo Kyung, Choung-Soo Kim, and Namkug Kim. Active learning for accuracy enhancement of semantic segmentation with cnn-corrected label curations: Evaluation on kidney segmentation in abdominal ct. *Scientific Reports*, 10(1): 366, 2020.

Andreas Kirsch, Joost R. van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *NeurIPS*, 2019.

Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. Cereals - cost-effective region-based active learning for semantic segmentation. In *BMVC*, page 121, 2018.

Amin Madani, Babak Namazi, Maria S. Altieri, Daniel A. Hashimoto, Ángela María Morales Rivera, Philip H. Pucher, Allison Navarrete-Welton, Ganesh Sankaranarayanan, L. Michael Brunt, Allan Okrainec, and Adnan A. Alseidi. Artificial intelligence for intraoperative guidance: Using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. *Annals of Surgery*, 2020.

Lena Maier-Hein, Swaroop S. Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, Makoto Hashizume, Darko Katic, Hannes Kenngott, Michael Kranzfelder, Anand Malpani, Keno März, Thomas Neumuth, Nicolas Padoy, Carla Pugh, Nicolai Schoch, Danail Stoyanov, Russell Taylor, Martin Wagner, Gregory D. Hager, and Pierre Jannin. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9): 691–696, 2017.

Salman Maqbool, Aqsa Riaz, Hasan Sajid, and Osman Hasan. m2caiseg: Semantic segmentation of laparoscopic images using convolutional neural networks. *ArXiv*, abs/2008.10134, 2020.

Pietro Mascagni, Deepak Alapatt, Takeshi Urade, Armine Vardazaryan, Didier Mutter, Jacques Marescaux, Guido Costamagna, Bernard Dallemagne, and Nicolas Padoy. A computer vision platform to automatically locate critical events in surgical videos. *Annals of Surgery*, 274(1), 2021.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Yawar Siddiqui, Julien Valentin, and Matthias NieBner. Viewal: Active learning with viewpoint entropy for semantic segmentation. pages 9430–9440, 06 2020.

Jamshid Sourati, Ali Gholipour, Jennifer G. Dy, Sila Kurugol, and Simon K. Warfield. Active deep learning with fisher information for patch-wise semantic segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, page 83–91, 2018.

Andru Putra Twinanda, S. Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36:86–97, 2017.

Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *ACCV*, 2020.

Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne, editors, *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, pages 399–407, Cham, 2017. Springer International Publishing.

Appendix A.

Simulation details: In this work, we used the *CholecSeg8k* dataset as a training bed for our method. Figure 5 illustrates samples from the *CholecSeg8k* dataset. All frames in this dataset have already been labeled, but we treated some initial pool of these images as the "labeled training data pool $\mathcal{D}_L$" and considered the rest as "unlabeled data pool $\mathcal{D}_U$". To simulate an external oracle or human annotator labeling the selected data by active learning, we simply included the labels of the selected images in the training data pool for that round of active learning. In practice, this is the equivalent of an expert clinician segmenting images given by the chosen active learning framework at each round of evaluation.

Appendix B.

Uncertainty based methods: We detail a few established active learning algorithms initially intended for the classification setting modified for the semantic segmentation setting. These methods select the top B images from the unlabeled data pool $\mathcal{D}_U$ ordered by various measures of total uncertainty $U_{\mathbf{X}}$:

For Max Entropy, batches are selected by measuring predictive entropy of each sample:

$$\{X_1^*, \ldots, X_B^*\} = \underset{\{X^1,\ldots,X^B\}\in\mathcal{D}_U}{\operatorname{argmax}} \sum_B U_{\mathbf{X}^b} = \underset{\{X^1,\ldots,X^B\}\in\mathcal{D}_U}{\operatorname{argmax}} \sum_B \sum_{ij} U_{\mathbf{x}_{ij}^b} \tag{7}$$

$$= \underset{\{X^1,\ldots,X^B\}\in\mathcal{D}_U}{\operatorname{argmax}} \sum_B \sum_{ij} \left(-\sum_{k=1}^{K} p_{ijk}^b \log(p_{ijk}^b) \right) \tag{8}$$

For Margins sampling, uncertainty is measured by the difference between the top two predicted classes for a sample. A smaller margin indicates higher predictive uncertainty:

$$\{X_1^*, \ldots, X_B^*\} = \underset{\{X^1,\ldots,X^B\}\in\mathcal{D}_U}{\operatorname{argmin}} \sum_B U_{\mathbf{X}^b} = \underset{\{X^1,\ldots,X^B\}\in\mathcal{D}_U}{\operatorname{argmin}} \sum_B \sum_{ij} U_{\mathbf{x}_{ij}^b} \tag{9}$$

$$= \underset{\{X^1,\ldots,X^B\}\in\mathcal{D}_U}{\operatorname{argmin}} \sum_B \sum_{ij} \left(\max_1(\mathbf{p}_{ij}^b) - \max_2(\mathbf{p}_{ij}^b) \right) \tag{10}$$

For Least confidence, uncertainty is inversely correlated with the magnitude of the highest predicted class for a sample. If the maximum predicted class probability is high, then the model is highly certain that the sample is of that class:

$$\{X_1^*, \ldots, X_B^*\} = \underset{\{X^1,\ldots,X^B\}\in\mathcal{D}_U}{\operatorname{argmin}} \sum_B U_{\mathbf{X}^b} = \underset{\{X^1,\ldots,X^B\}\in\mathcal{D}_U}{\operatorname{argmin}} \sum_B \sum_{ij} U_{\mathbf{x}_{ij}^b} \tag{11}$$

$$= \underset{\{X^1,\ldots,X^B\}\in\mathcal{D}_U}{\operatorname{argmin}} \sum_B \sum_{ij} \left(\max_1(\mathbf{p}_{ij}^b) \right) \tag{12}$$

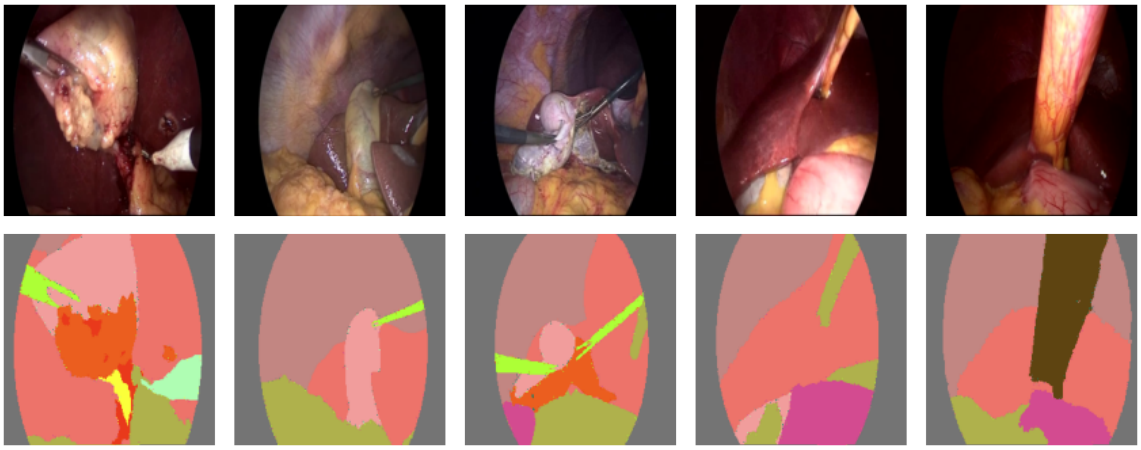

Figure 5: Qualitative examples of laparoscopic surgery video data with ground truths.