Correlation-aware active learning for surgery video segmentation

Fei Wu
AIMI, University of Bern
fei.wu@unibe.ch

Pablo Marquez-Neila
AIMI, University of Bern
pablo.marquez@unibe.ch

Mingyi Zheng
Auris Health. Inc, JNJ
immingyizheng@gmail.com

Hedyeh Rafii-Tari
Auris Health. Inc, JNJ
hrafiita@its.jnj.com

Raphael Sznitman
AIMI, University of Bern
raphael.sznitman@unibe.ch

Abstract

Semantic segmentation is a complex task that relies heavily on large amounts of annotated image data. However, annotating such data can be time-consuming and resource-intensive, especially in the medical domain. Active Learning (AL) is a popular approach that can help to reduce this burden by iteratively selecting images for annotation to improve the model performance. In the case of video data, it is important to consider the model uncertainty and the temporal nature of the sequences when selecting images for annotation. This work proposes a novel AL strategy for surgery video segmentation, COWAL, *COrrelation-aWare Active Learning. Our approach involves projecting images into a latent space that has been fine-tuned using contrastive learning and then selecting a fixed number of representative images from local clusters of video frames. We demonstrate the effectiveness of this approach on two video datasets of surgical instruments and three real-world video datasets. The datasets and code will be made publicly available upon receiving necessary approvals.*

1. Introduction

Minimally invasive surgical robotic systems have seen widespread adoption for surgical procedures. In this context, endoscopic video feeds allow new possibilities to enhance, augment and even partially automate specific tasks. This includes clinical decision support [28], case review [45], or even intra-operational fusion [8]. However, a key necessity for these applications is the ability to segment surgical instruments [2, 27, 46]. To this end, promising performances for binary segmentation have been shown, but these methods suffer from the need for large amounts of manually annotated data to train subsequent models.

To alleviate the segmentation annotation burden, Active Learning (AL) is a well-established learning strategy [5, 9, 36, 37, 56, 59]. Using a large set of unlabeled data, AL iteratively optimizes which data points should be annotated by an oracle (*i.e.*, expert) to improve the model performance most efficiently. To date, it has been widely applied to different clinical applications such as echocardiography view classification [58], diabetic retinopathy detection [49], or surgical phases recognition [43].

Video sequences pose unique challenges for AL methods due to the high correlation between video frames. An example of this can be seen in Fig. 2 (rows 2, 4, 5, 6 and 7), where consecutive frames exhibit high visual similarity. Annotating more than one of these redundant frames would bring just a marginal information gain to the segmentation model and, considering the high cost of segmentation annotations, should be avoided by an effective AL method. However, while typical segmentation AL methods such as entropy sampling incorporate some implicit or explicit mechanism to prevent redundant samples from being annotated, they are designed under the assumption that image samples are independent and identically distributed, which is not satisfied in the case of video sequences and necessarily leads to suboptimal sample choices. This is especially problematic in deep learning AL methods, where multiple samples sharing similar properties are simultaneously selected at each step. Therefore, video AL methods must be designed to account for the highly correlated nature of video sequences explicitly.

Even though many works have applied AL for videos [6, 7, 12, 18, 23, 52] and semantic segmentation [17, 47, 48, 54, 56, 59], only a few have studied AL methods for semantic segmentation on video datasets [35, 39]. Peng *et al*. [39] used the EndoVis 2017 dataset [3], which contains surgery videos, but they curated the dataset to remove similar frames, hence the challenge within video datasets, and used a sampling strategy based on model uncertainty. Mittai *et al*. [35] compared existing AL methods for video datasets. They pointed out the inefficiency of uncertainty-

based methods on videos because these methods do not consider the information redundancy in video datasets. They showed that a diversity-based method Core-Set [41], is a better choice of sampling strategy for video datasets. We validate their finding and further show that our approach, combining uncertainty and diversity, can yield better performance for surgery videos.

We propose *COWAL*, an AL strategy for video segmentation. Given that it is computationally advantageous to sample multiple images when iteratively training NNs, we hypothesize that the selection of images should consider both the model uncertainty and the temporal nature of video sequences. We propose to project images in a latent space, fine-tuned by contrastive learning, and then select a fixed number of images at each iteration representative of local clusters of video frames. We show experimentally that our approach is superior to standard AL selection strategies.

2. Related Works

We present here some of the relevant related works. These consist of AL methods for video applications, methods for semantic segmentation, and general AL methods.

2.1. Semantic Segmentation

Semantic segmentation is the task of classifying each pixel of an image into a defined category. Several AL methods for semantic segmentation have been proposed in the past. Broadly, these can be categorized as methods that sample images [17, 39, 48, 56, 59] or sample image regions [10, 11, 15, 22, 32, 33, 40, 47, 50, 53, 54]. While sampling image regions allows for finer-grained queries (*i.e*. labeling of a group of adjacent pixels in a frame), it often comes at the expense of more complex AL methods with higher performance variance. Instead, this work focuses on image-level sampling, allowing for more scalable AL, and additional annotation costs are marginal in time and effort compared to neural network training time.

In the category of methods that sample images, Yang *et al*. [56] first considers a subset S of samples from the unlabeled pool with the highest entropy, usually two times the number of the query budget. They proceed to iteratively select samples from this subset that best cover the unlabeled data distribution in an embedding space. Unlike them, we also take into account the embedding of labeled data and group all embeddings into distinct clusters before sampling the image with the highest uncertainty from each cluster.

Sinha *et al*. [48] use adversarial training between a VAE [30] and a discriminator to generate image embedding using the VAE and having the discriminator learn the difference between embeddings of labeled images and unlabeled images at every AL step. Once the discriminator is trained, they select unlabeled samples that are most different from labeled samples according to the discriminator. Instead of training a VAE at every AL step, we train an embedding model once using contrastive learning to learn image similarity. We then cluster unlabeled images with visually similar labeled images and sample unlabeled images that are not grouped with labeled images.

2.2. Video Datasets

Due to the vast amounts of data available in video content, AL learning for video has drawn research interests in recent years [6, 7, 12, 18, 23, 35, 39, 52]. These methods have focused on a wide range of tasks, such as video caption learning [12], surgical phase classification [44], video object detection [7], video tracking [6, 52], video object segmentation [23], and video semantic segmentation [35].

While Griffin and Corso [23] developed a method for video object segmentation, this task only requires the annotator to label one frame per video, and a model will learn to track the segmented object of this one frame throughout the video. In contrast, our method handles semantic segmentation and requires the sampling of several frames. Thus the method developed by Griffin and Corso [23] can not be directly applied in our case.

The work of Mittal *et al*. [35] is, however, relevant to ours. Similar to our findings, they have pointed out the inefficiency of the uncertainty-based sampling strategy for datasets with redundant images. They applied AL methods to the A2D2 [21] and Cityscapes [16] datasets. Even though Cityscapes [16] comes from video sequences, it is more akin to image datasets because of the curation it went through to remove redundant frames. We thus focus on their result for the A2D2 dataset [21], and their finding is that a diversity-based sampling strategy Core-Set [41] performs better than uncertainty-based methods, which aligns with our experiments. They have nevertheless not tested sampling methods that combine both uncertainty and diversity. We show in this paper a similar study focusing on surgery videos. However, unlike Mittal *et al*. [35], who compared existing AL methods for video segmentation, we also propose our own sampling strategy based on uncertainty-diversity.

2.3. Active Learning

A recent survey on AL can be found in the work of Tharwat *et al*. [51] and a survey on AL applied to the medical domain [9]. AL methods can be categorized into three groups: uncertainty-based [19, 20, 26, 38, 42], diversity-based [41, 48], and methods that combine both uncertainty and diversity [4, 31, 34, 50, 56]. Our approach falls into the last category and proposes to select samples based on their diversity and the uncertainty of the model.

In the same category as our method, Ash *et al*. [4] compute the gradient embeddings of the model classification predictions and use the magnitude of these embeddings

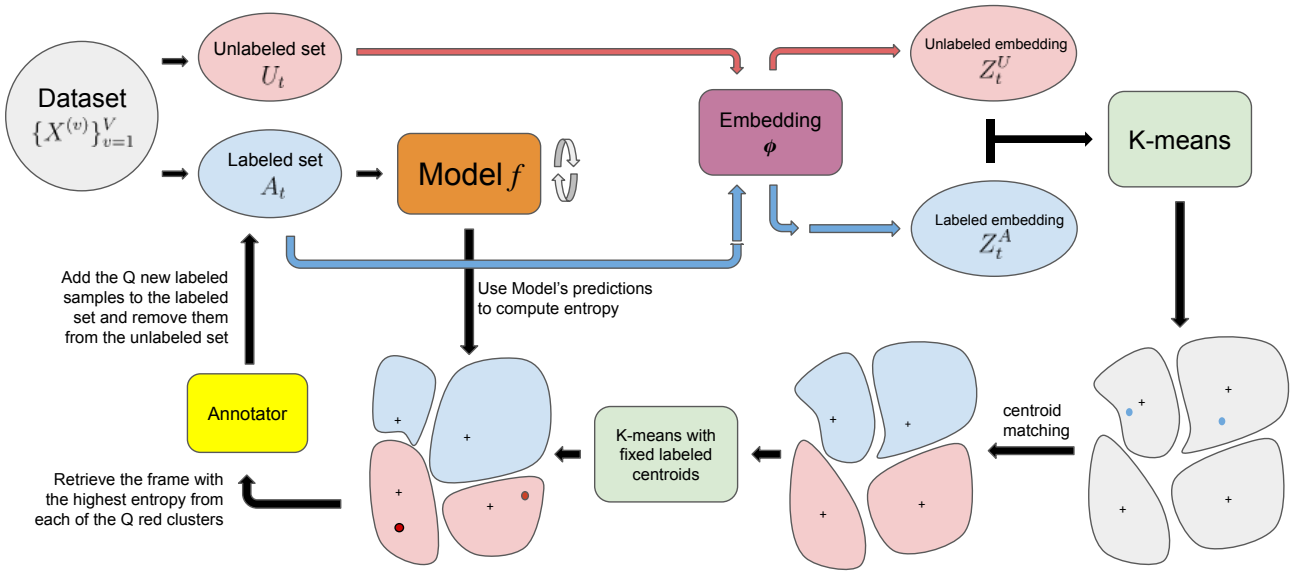

Figure 1. An overview of our approach. A subset of the dataset is labeled to train a model. Once the model has converged, an embedding function projects all the images in a latent space. K-means is then used to create clusters of similar images. We assign all labeled frames to the closest centroids and fix the centroids to those. K-means is applied again over the remaining centroids. Our strategy selects the samples with the highest model entropy per cluster for annotation.

to assess the uncertainty of the corresponding predictions. They proceed to sample inputs whose gradient embedding is far from each other using the k-means++ initialization method. The gradient embedding is obtained by multiplying the predicted probability of the image with a hidden layer embedding, a vector. In the case of segmentation, we have a 2D map of probabilities instead of a single scalar. Computing the gradient embedding by multiplying the probability of each pixel with the embedding and then stacking them would yield gigantic vectors, hence it is not straightforward to adapt their method for segmentation tasks.

Margatina *et al*. [34] get the embeddings for input text data and sample inputs whose classification predictions differ despite having similar embeddings. Their method assumes the prediction is a probability vector and can not be applied directly to segmentation tasks.

Other methods, such as [31, 38] also rely on the classification nature of the task, and it is not straightforward to adapt them for segmentation. We thus compare our approach to the following uncertainty-based methods [19, 20, 42], diversity-based methods [41, 48] and uncertainty-diversity methods [56].

3. Method

A video segmentation model is a function $f : \mathcal{X} \rightarrow \mathcal{Y}$ that receives an image, or video frame, $\mathbf{x} \in \mathcal{X}$ and produces a label map $\mathbf{y} = f(\mathbf{x}) \in \mathcal{Y}$. Training such a model requires pairs $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$ of video frames and their corresponding annotated label maps. While obtaining video frames is usually inexpensive, producing manual annotations is time-consuming and effort-intensive. Active learning is a technique that reduces the number of required annotations by iteratively selecting the most informative frames for annotation. At each step t, a subset of the unlabeled frames is selected according to an *annotation strategy function* $\pi : 2^{\mathcal{X}} \times 2^{\mathcal{X} \times \mathcal{Y}} \rightarrow 2^{\mathcal{X}}$ that receives the set of unlabeled frames $\mathcal{U}_t$ and the set of labeled frames $\mathcal{A}_t$ and suggests a subset of frames $\pi(\mathcal{U}_t, \mathcal{A}_t) \subset \mathcal{U}_t$ that should be annotated. After annotation, these frames are moved to the set of labeled frames for the next step $\mathcal{A}_{t+1} \subset \mathcal{X} \times \mathcal{Y}$. The model is then trained with $\mathcal{A}_{t+1}$, and the whole process is repeated in the next step.

In this work, we focus on the problem of active learning with video sequences. Given a collection of unlabelled videos $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$, where each video is a sequence of frames $\mathbf{X}^{(v)} = (\mathbf{x}_1^{(v)}, \ldots, \mathbf{x}_{F_v}^{(v)})$, our goal is to find a frame annotation strategy π that maximizes the performance of the segmentation model f for a fixed annotation budget. This problem is challenging since video frames are not independent and identically distributed (*i.i.d.*), a default assumption in most active learning methods. While it is possible to treat the set of frames as *i.i.d.* samples, in practice, video frames exhibit strong temporal dependencies that we can exploit in the design of π.

Figure 2. First 2 rows show examples of images obtained from the MONARCH™ Platform. Row 1 shows different instruments (various sheaths, needle, brush, forceps, REBUS probe) in the dataset and their corresponding segmentation masks. Row 2 shows partial image sequences at 2 fps. Row 3 and 4 show examples of images from the EndoVis [2] dataset. Row 3 shows different tools (instrument shaft, instrument clasper, instrument wrist, thread, clamps, needle, suction instrument, and ultra-sound probe) in the dataset and their corresponding segmentation masks. Row 4 partial image sequence at 1 fps. Row 5 shows partial image sequences of the A2D2 [21] dataset. Row 6 and 7 show image sequences of the Skateboard and Parrot datasets [55] at 6 fps.

3.1. Annotation strategy

Our strategy is to select unlabeled frames representing highly correlated video segments while avoiding selecting redundant frames similar to those already in the labeled dataset. At each step t, the policy first projects all frames in $\mathcal{A}_t$ and $\mathcal{U}_t$ to a representation space $\mathcal{Z}$ with an embedding function $\phi : \mathcal{X} \rightarrow \mathcal{Z}$. K-means are applied in this space with $K = |\mathcal{A}_t| + Q$, where $|\cdot|$ is the cardinal of a set, and Q is a hyperparameter indicating the desired number of selected frames at each step. This process finds a set $\{\mathbf{k}^{(i)}\}_{i=1}^{K}$ of representative centroids in the embedding space. We then look for centroids similar to the frames in $\mathcal{A}_t$ in terms of distance in the embedding space.

We design a matching algorithm (see Appendix) to find the matching $\{(\mathbf{a}^{(i)}, \mathbf{k}^{(m_i)})\}_{i=1}^{|\mathcal{A}_t|}$ between the centroids $\mathbf{k}$ and the embeddings $\mathbf{a}$ of the elements of $\mathcal{A}_t$. The distance between a centroid $\mathbf{k}$ and $\mathcal{A}_t$ is defined as $\min_{\mathbf{a}\in\mathbf{A}_t} \|\mathbf{a}-\mathbf{k}\|$. We first assign the centroids $\mathbf{k}$ with the smallest distances to $\mathcal{A}_t$ with the corresponding embedding $\mathbf{a}$ used in the distance calculation. Since we have more centroids than frames in $\mathcal{A}_t$, this matching ensures that Q centroids with the highest distances to $\mathcal{A}_t$ are left out. Each matched centroid $\mathbf{k}^{(m_i)}$ is substituted by its corresponding vector $\mathbf{a}^{(i)}$.

A second round of k-means is applied to update the remaining Q unmatched centroids while keeping the matched centroids fixed to their new values $\mathbf{a}^{(i)}$. This process ensures that the new Q centroids will represent video segments not already covered by the labeled samples since labeled samples belong to the clusters of the fixed centroids. Finally, the selected samples by our strategy are those with the highest entropy [2] from each cluster of the Q unmatched centroids. The entropy of a frame is computed as the sum of pixel entropies, which are computed in a standard way by applying the entropy formula to the output probabilities of the segmentation model. The probabilities are given after the softmax or sigmoid operation.

Note that just running k-means once and fixing $|\mathcal{A}_t|$ centroids to the corresponding embeddings of labeled frames is highly dependent on the initialization of the remaining Q centroids, and we found experimentally to lead to suboptimal solutions, as the fixed centroids prevent the others from moving and adequately covering the embedded vectors.

Our strategy π assumes that the distance between embedding vectors represents the level of correlation between frames. In particular, the distance between highly correlated frames should be smaller than between independent frames. We design our embedding function accordingly.

3.2. Representation learning

The embedding function $\phi : \mathcal{X} \rightarrow \mathcal{Z}$ is modeled as a ResNet-34 [25] trained in an unsupervised way with SimCLR [14] once before AL starts. In particular, SimCLR minimizes the contrastive loss

$$\ell_{i,j} = -\log \frac{\exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp\left(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau\right)} \quad (1)$$

where sim is the cosine similarity $\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^T \cdot \mathbf{v}}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|}$, $\mathbf{z}_i = \phi(\mathbf{x}_i)$. The loss is minimized for all pairs of frames augmented from the same image.

4. Experiments

4.1. Dataset

MONARCH. We collected and created a video dataset using the MONARCH™ Platform, a robotic-assisted bronchoscopy system indicated for diagnostic and therapeutic procedures within the lung. The dataset centers on the biopsy phase, encompassing the insertion of diverse instruments such as REBUS, needles, forceps, and brushes. It contains 11 videos of bronchoscopy procedures with segmentation masks on the biopsy tools acquired following

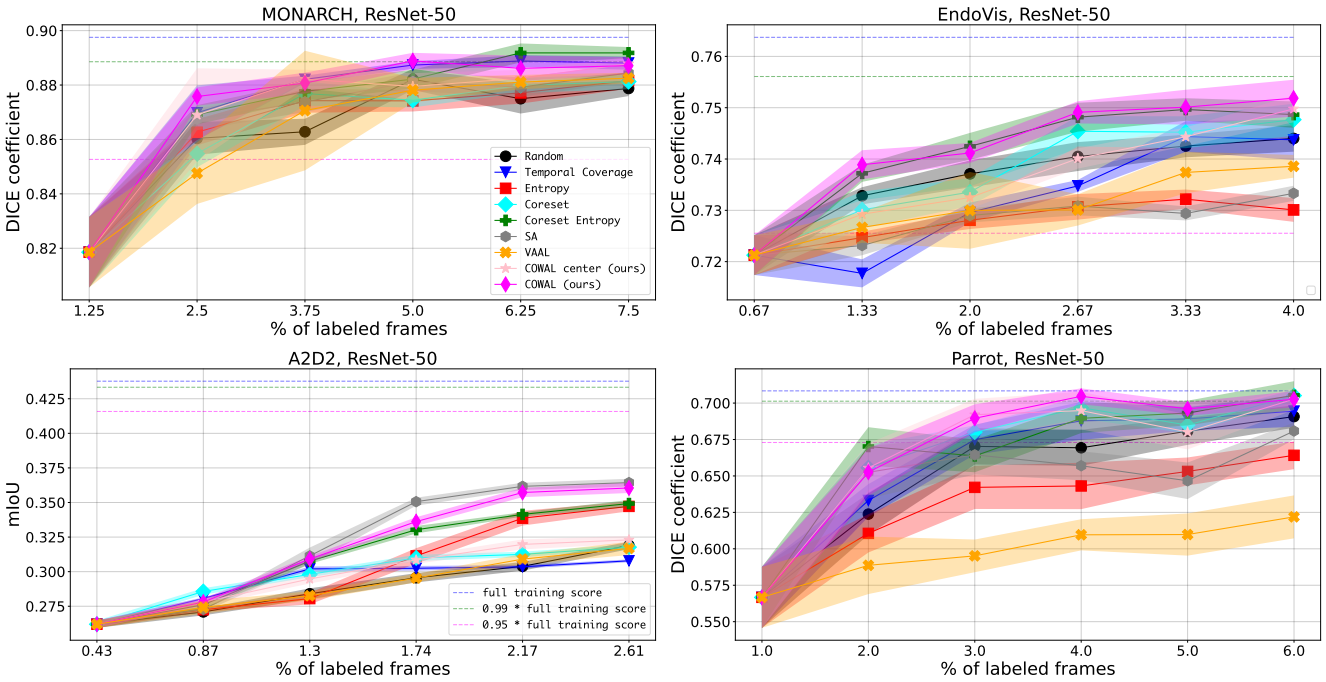

Figure 3. Comparison of different sampling strategies. Values are averaged over 10 different training-validation splits, with error bars indicating one standard deviation. Plots for Skateboard and ResNet101 backbones are available in Appendix A.

HIPAA requirements. These videos were subdivided into 56 segments, collectively amounting to 2883 frames. Segments devoid of visible biopsy tools were excluded from the experiments.

The videos were recorded at 25 fps with a resolution of 220×220, and the segmentation annotation were done at 2 fps with a labeled frame every 12 frames. Examples of the different tools and their segmentation mask are displayed in Fig. 2. In this dataset, only one tool is visible per frame. We separate the 56 videos into 30 videos for training ($\approx 1'600$ frames) and 26 videos for testing ($\approx 1'300$ frames). For our experiments, the dataset is simplified from a multi-tool segmentation task to a binary segmentation task with biopsy tools versus background annotations.

EndoVis [2]. We also evaluate our method on the EndoVis 2018 Robotic Scene Segmentation dataset [2]. It is a dataset of surgery videos with segmentation masks on surgery tools and human body organs. A total of 19 videos are divided into 15 training videos ($\approx$ 2250 frames) and 4 test videos ($\approx$ 1000 frames). Each video came from a single porcine training procedure. Images from the left eye in the stereo pair are used for training. The annotation procedure is done at 1 fps with non-anatomical classes such as instrument shaft, instrument wrist, instrument clasper, threads, clamps, suturing needle, suction instrument, and ultra-sound probe. An example of each tool is shown in Fig. 2 row 3. Like the MONARCH dataset, we simplified the annotations with an anatomical vs. non-anatomical binary mask.

A2D2 [21] is a large-scale driving dataset consisting of 41277 annotated images with a resolution of 1208×1920 from 23 sequences. It covers an urban setting from highways, country roads, and three cities. It contains labels for 38 categories. For our experiments, we map them to the 19 classes of Cityscapes. Following the procedure of [35], we extract 60 video segments from the original A2D2 dataset. Each segment contains 44 consecutive frames, totaling 2640 frames. The video *'20180925_112730'* with 993 frames is used as the test set. The annotation rate is irregular, varying between 10 to 300 frames per annotation.

YouTube-VOS [55] is the most extensive video segmentation dataset with more than 5000 videos and 90 classes. We selected only videos showing skateboarders or parrots to build two binary segmentation datasets. The Skateboard dataset contains 24 videos ($\approx$ 700 frames) for training and 10 videos ($\approx$ 250 frames) for testing. The Parrot dataset contains 46 videos ($\approx$ 1500 frames) for training and 10 videos ($\approx$ 500 frames) for testing.

4.2. Baselines

We evaluate *COWAL* against the following baselines methods, *Random*, *Entropy* [42], *MC Dropout* [19], *BALD* [20], *Suggestive Annotation* [56], *CoreSet* [41],

Sampling Method	Backbone	MONARCH	EndoVis	A2D2	Skateboard	Parrot	Avg across datasets
Random	ResNet50	0.804	0.804	0.550	0.790	0.770	0.744
Temporal Coverage	ResNet50	<u>0.813</u>	0.799	0.561	0.746	0.780	0.740
Entropy [42]	ResNet50	0.805	0.795	0.574	0.768	0.745	0.737
CoreSet [41]	ResNet50	0.805	0.805	0.570	0.789	0.789	0.752
CoreSet x Entropy	ResNet50	<u>0.813</u>	<u>0.810</u>	0.594	0.781	0.789	<u>0.757</u>
SA [56]	ResNet50	0.808	0.794	**0.613**	0.769	0.763	0.749
VAAL [48]	ResNet50	0.804	0.797	0.552	0.740	0.705	0.720
COWAL center (ours)	ResNet50	0.811	0.803	0.569	**0.799**	<u>0.790</u>	0.754
COWAL (ours)	ResNet50	**0.814**	**0.811**	<u>0.606</u>	<u>0.793</u>	**0.795**	**0.764**
Random	ResNet101*	0.815	0.809	0.584	<u>0.804</u>	0.818	0.766
Temporal Coverage	ResNet101*	<u>0.817</u>	0.803	0.579	0.776	<u>0.846</u>	0.764
Entropy [42]	ResNet101*	0.810	0.799	0.589	0.785	0.781	0.753
CoreSet [41]	ResNet101*	0.815	0.811	0.590	0.802	0.831	0.770
CoreSet x Entropy	ResNet101*	<u>0.817</u>	<u>0.812</u>	0.616	0.787	0.835	<u>0.773</u>
SA [56]	ResNet101*	0.814	0.801	**0.632**	0.790	0.797	0.767
VAAL [48]	ResNet101*	0.811	0.802	0.586	0.763	0.759	0.744
MC Dropout [19]	ResNet101*	0.810	0.800	.	.	.	.
BALD [20]	ResNet101*	0.805	0.806	.	.	.	.
COWAL center (ours)	ResNet101*	0.812	0.810	0.581	**0.806**	0.835	0.769
COWAL (ours)	ResNet101*	**0.819**	**0.816**	<u>0.621</u>	<u>0.804</u>	**0.848**	**0.782**

Table 1. AuALC of all sampling methods across all datasets. Metrics are computed at the end of the active learning steps. The $*$ in ResNet101* indicates that it is a Bayesian model which allows the use of sampling methods such as MC Dropout [19] or BALD [20]. Best scores are in bold and second best are underlined.

VAAL [48] as well as some custom designed baselines:
- ***Temporal Coverage*** selects samples by prioritizing videos with fewer labeled frames, and within a video, select the temporally most distant frame to currently labeled frames.
- ***CoreSet x Entropy*** computes the distance of an unlabeled sample to the labeled set as in *CoreSet* [41] and scale these distances by the entropy values [42], selecting the samples with the highest scaled distances. We use the same embedding ϕ as *COWAL*.
- ***COWAL center***, we also compare *COWAL* with an ablated version of itself. Instead of selecting the frame with the highest entropy per cluster, we select the frame closest to the centroid to maximize sample diversity.

4.3. Implementation details

We follow the same experimental setup for all the baselines. The MONARCH input images are 220×220, the EndoVis images are 224×224, the A2D2 images are 270×480, and the YouTube-VOS images are 256×448. For data augmentation, we take crops with random scale factors in the range $(0.85, 1)$ of the original image and with random aspect ratios in the range $(\frac{3}{4}, \frac{4}{3})$. All crops are rescaled back to the size of the original image, followed by a random horizontal flipping with a probability of 0.5.

The set of labeled frames $\mathcal{A}_1$ is initialized with the middle frame of 10 training videos and their corresponding annotations. The complementary set of unlabeled frames $\mathcal{U}_1$ contains the remaining frames from the training data. At each step t, the segmentation model f is trained with $\mathcal{A}_t$ until convergence of the DICE score on the validation set. We run each baseline 10 times, randomly splitting the training videos into training and validation sets with a proportion of $2:1$ for the EndoVis dataset, $1:1$ for the MONARCH, $9:1$ for A2D2, $2:1$ for Parrot, and $3:1$ for Skateboard. The reason for a $1:1$ split for MONARCH is that the videos from this dataset are usually shorter and contain many redundant frames. Hence, having many videos in the validation set ensures it contains diversified images. The segmentation model f is a DeepLabV3 [13] for which we have two different versions. The first version is from the official PyTorch/vision GitHub repository [1] and uses the ResNet-50 backbone. The second version, from the implementation of [47], uses a ResNet-101 backbone and a deeper decoder with more dropout layers.

The training is done with a batch size of 4, a learning rate of 10^{-4}, and no weight decay using the Adam optimizer [29]. The patience is set to 20, and at every AL step, evaluation on the validation set is done after the number of iterations required to go through the whole training set. We apply Polyak averaging with $\alpha = 0.99$ for stable training evolution. At the first AL step $t = 1$, the model is initialized with ImageNet pre-trained weights and, for subsequent steps, using the weights obtained at the end of the previous step. The policy function chooses $Q = 10$ new samples

for annotation at each AL step. After convergence, model performance is evaluated with the DICE score on the test set.

We train the embedding function ϕ with contrastive learning on the train and validation sets of the respective dataset as described in Sect. 3.2. We used a learning rate of $3 \cdot 10^{-4}$, a batch size of 256, a weight decay of 10^{-2}, and trained for 2000 epochs using the Adam optimizer [29].

4.4. Evaluation protocol

We run each baseline 10 times, and perform 6 AL steps on each run. We compute the median of the DICE scores over the 10 runs for each AL step and plot the resulting AL curves of *median DICE* vs. *AL step* to compare the performances. In addition, we report the area under the AL curve (AuALC) as our evaluation metric. The AuALC is expressed as a fraction of the maximum possible area that a hypothetically perfect AL method would achieve if it could reach the same performance as training the model with the complete training dataset at every AL step.

5. Results

Fig. 3 shows the experimental results. The differences in performance among baselines are significant, as indicated by the standard errors.

COWAL outperformed all baselines for 3 datasets out of 5 and places second in the remaining 2 datasets (Table 1). Strategies that rely only on uncertainty, such as *Entropy* [42], *MC Dropout* [19], and *BALD* [20], performed poorly across all different settings in our experiments. These strategies have proven highly effective for image datasets [24, 57], but they are more likely to select redundant frames when querying batches of frames from video sequences, as shown in Fig. 4. For example, *Entropy* selected multiple consecutive frames of video 9. This happens because frames that are temporally close often have similar uncertainty scores, causing a concentration of high-uncertainty frames within temporal neighborhoods. In contrast, *COWAL*, by ensuring a degree of visual diversity among labeled frames, selected only one frame from the same video.

While ***Temporal Coverage*** performed well on MONARCH, it struggled during the first AL iterations on the EndoVis dataset. *Temporal Coverage* starts by selecting the frames from the beginning and end of each video sequence, which often lack relevant information. For example, in Fig. 4 Row 1, selected frames **v3-f0** and **v6-f0** contain minimal presence of surgery tools, negatively impacting performance in the initial AL iterations.

Suggestive Annotation [56] outperforms *COWAL* on the A2D2 dataset due to lower frame redundancy. *Suggestive Annotation* selects high-entropy samples and diversifies within them, while *COWAL* diversifies first and then identifies high-entropy frames. *Suggestive Annotation* generally prioritizes higher-entropy samples over *COWAL*, yet frame redundancy can limit its effectiveness. In A2D2, with fewer redundant high-entropy samples, *Suggestive Annotation* performs better. For similar reasons, *Entropy* [42] surpasses *Random* and other diversity-based methods exclusively in the A2D2 dataset. The selected *Suggestive Annotation* frames are listed in Appendix D.

Sampling Method	Backbone	MONARCH	EndoVis
CoreSet with TME [41]	ResNet50	0.718	0.795
CoreSet [41]	ResNet50	0.718	**0.805**
SA with TME [56]	ResNet50	**0.723**	**0.796**
SA [56]	ResNet50	0.722	0.794
COWAL with TME (ours)	ResNet50	0.721	0.803
COWAL (ours)	ResNet50	**0.727**	**0.811**
CoreSet with TME [41]	ResNet101*	0.814	0.805
CoreSet [41]	ResNet101*	**0.815**	**0.811**
SA with TME [56]	ResNet101*	0.813	**0.803**
SA [56]	ResNet101*	**0.814**	0.801
COWAL with TME (ours)	ResNet101*	0.817	0.812
COWAL (ours)	ResNet101*	**0.819**	**0.816**

Table 2. AuALC of embedding ablation across *CoreSet*, *Suggestive Annotation*, and *COWAL* sampling strategies. The embedding method of each approach uses either the Task Model Embedding (TME) or the embedding ϕ.

5.1. Embedding Model Ablation

We evaluate the impact of removing our contrastive embedding function ϕ and using the embedding defined by the downstream segmentation model f instead, as it is done in [41,56]. As shown in Table 2, the use of the embedding ϕ yields better results for *COWAL* and *CoreSet*, while the task model embedding is helpful for *Suggestive Annotation*. AL curves for this ablation are in Appendix B.

Suggestive Annotation performs diversity sampling on a subset S of the unlabeled samples, unlike the other two methods, which use the whole unlabeled set. The subset S has high entropy frames, which for video datasets means similar frames, and we argue that the quality of the embedding method is less relevant when applied to a pool of similar images hence explaining the different behavior of *Suggestive Annotation* compared to the other two methods.

5.2. Budget Size

We compare sampling strategies with an increase selection size ($Q = 50$) for the MONARCH-ResNet101 setting in Fig. 5. *COWAL* continues to outperform the other baselines. Among them, uncertainty-driven approaches, such as *Entropy* [42] and *BALD* [20], exhibit noticeably lower performance.

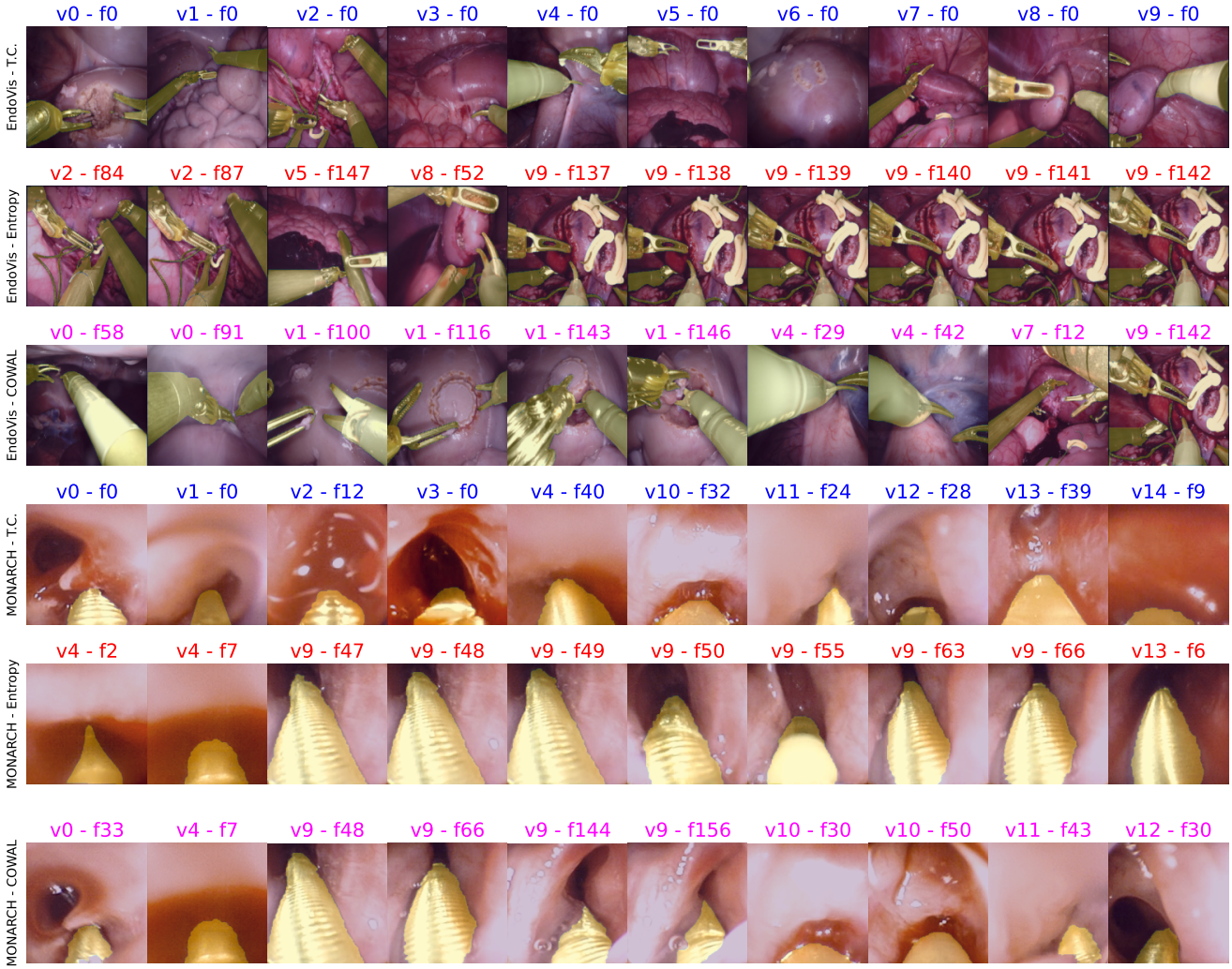

Figure 4. Selected frames at the first iteration. Video and frame numbers are indicated on top of each image.

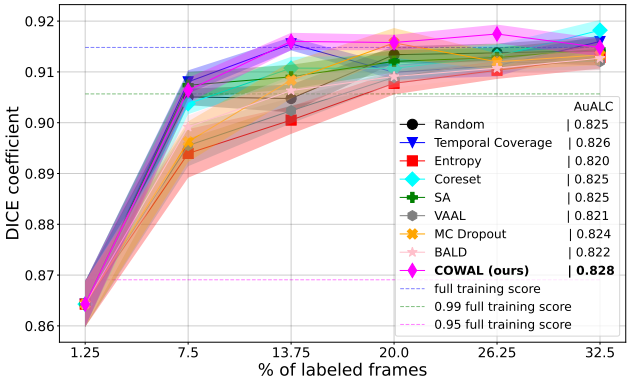

Figure 5. Comparison of different sampling strategies with a budget of $Q = 50$ instead of 10. AuALC is displayed on the side.

6. Conclusion

We present a novel AL approach for video segmentation that aims to sample images that are diverse from previously selected frames, and that are uncertain according to the model. We select maximal entropy frames for each cluster, yielded by modified iterated k-means that enforce previously selected frames to differ from new ones. We experimentally show the effectiveness of our approach against different AL selection schemes, whereby our approach does consistently better. We visually show that our approach indeed provides a good strategy to diversify labeled images, by still selecting informative samples.

Acknowledgement

This work is partially sponsored by Auris Health, Inc. (part of Johnson & Johnson MedTech).

Appendices

Appendix A. Additional plots

Figure 6 shows the results for the skateboard dataset using the ResNet-50 backbone, along with the results for all datasets using the ResNet-101 backbone.

Figure 7 shows additional comparisons between our contrastive embedding method and the standard task model embedding.

Appendix B. Centroid matching algoritm

Algorithm 1 describes the procedure to find a matching between the embedding of labeled frames and the k-means centroids, as required by *COWAL*.

Appendix C. Additional sampled frames

8 shows additional sampled frames by the *Temporal Coverage*, *Entropy*, *CoreSet*, and *COWAL* methods for the MONARCH dataset

9 shows additional sampled frames by the *Temporal Coverage*, *Entropy*, *CoreSet*, and *COWAL* methods for the EndoVis dataset.

10 shows additional sampled frames by the *Temporal Coverage*, *Entropy*, *CoreSet*, *Suggestive Annotation*, and *COWAL* methods for the A2D2 dataset.

11 shows additional sampled frames by the *Temporal Coverage*, *Entropy*, *CoreSet*, and *COWAL* methods for the Skateboard dataset.

12 shows additional sampled frames by the *Temporal Coverage*, *Entropy*, *CoreSet*, and *COWAL* methods for the Parrot dataset.

Algorithm 1 Centroid Matching

Require: Embedding of labeled frames $\{\mathbf{a}_i\}_{i=1}^{|\mathcal{A}|}$, K-means centroids $\{\mathbf{k}_j\}_{j=1}^{K}$ with $K = |\mathcal{A}| + Q$

$d[i,j] = \|\mathbf{a}_i - \mathbf{k}_j\|$ ▷ Build matrix of pair-wise distances

visit_order $\leftarrow$ argsort$_i \min_j d[i,j]$ ▷ Labeled frames will be visited according to their distance to the closest centroid.

$\mathcal{M} \leftarrow \emptyset$ ▷ Set of matches

assigned $\leftarrow \emptyset$ ▷ Set of assigned centroids

▷ For each frame in the given order...

for $i' \in$ visit_order **do**

  ▷ Visit the centroids sorted by distance to i'

  for $j' \in$ argsort$_j d[i',j]$ **do**

    ▷ Assign the first centroid not yet assigned

    if $j' \notin$ assigned **then**

      assigned $\leftarrow$ assigned $\cup \{j'\}$

      $\mathcal{M} \leftarrow \mathcal{M} \cup \{(i',j')\}$

      break

    end if

  end for

end for

▷ Return the set of matches

return $\mathcal{M}$

▷ A labeled frame i matches with centroid m_i if $(i, m_i) \in \mathcal{M}$

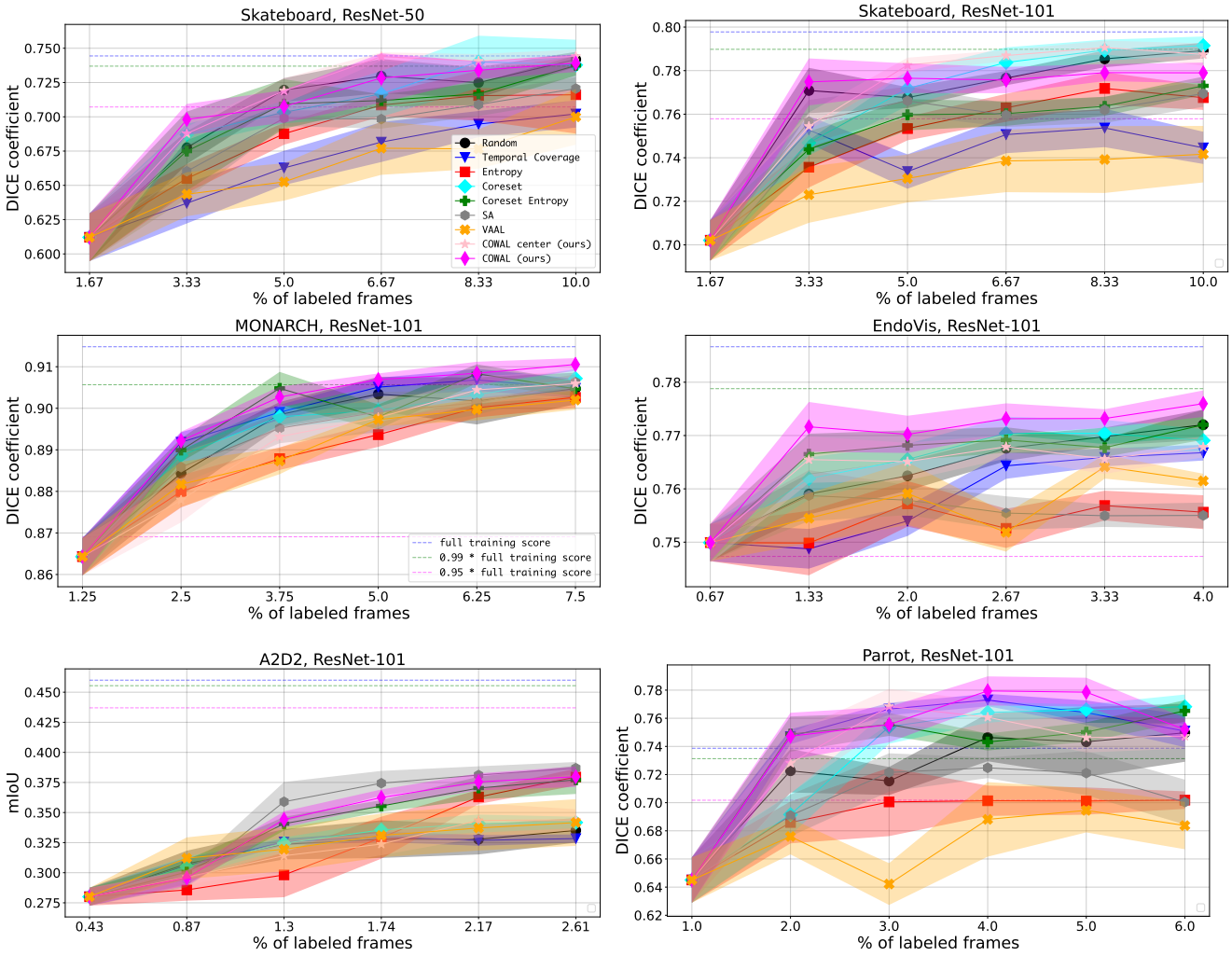

Figure 6. Appendix A. Comparison of different sampling strategies. Values are averaged over ten different training-validation splits, with error bars indicating one standard deviation.

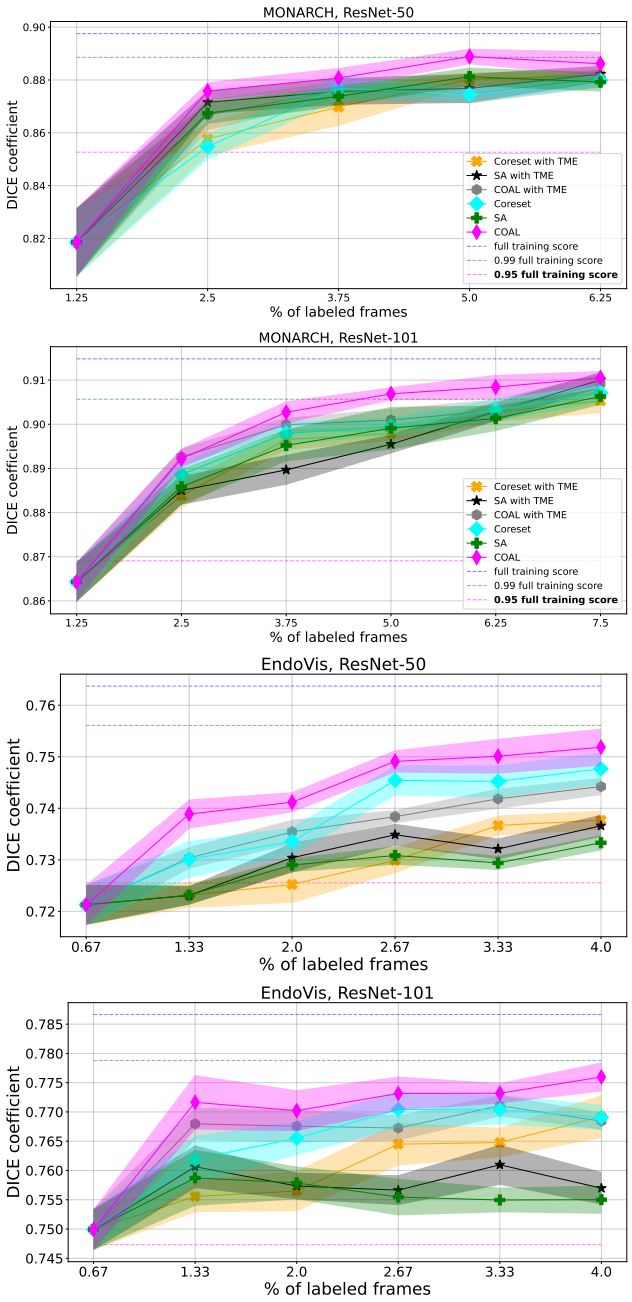

Figure 7. Comparison of *COWAL*, *CoreSet*, and *Suggestive Annotation* with two different embedding methods: Task Model Embedding (TME) or our contrastive embedding ϕ. Experiments are conducted on the MONARCH and EndoVis datasets with the ResNet-50 and ResNet-101 backbones. Evaluations are repeated for 10 different training-validation splits with error bars indicating one standard deviations.

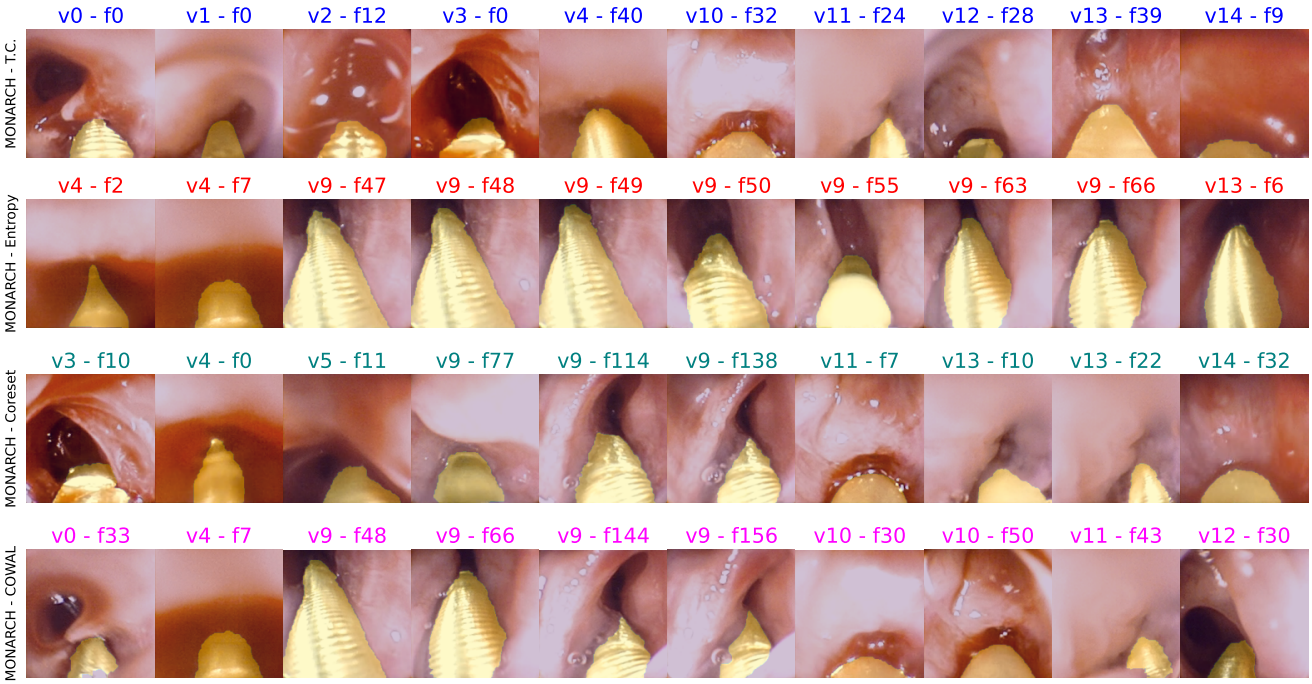

Figure 8. Selected frames at the first iteration on the MONARCH dataset. Video and frame numbers are indicated on top of each image.

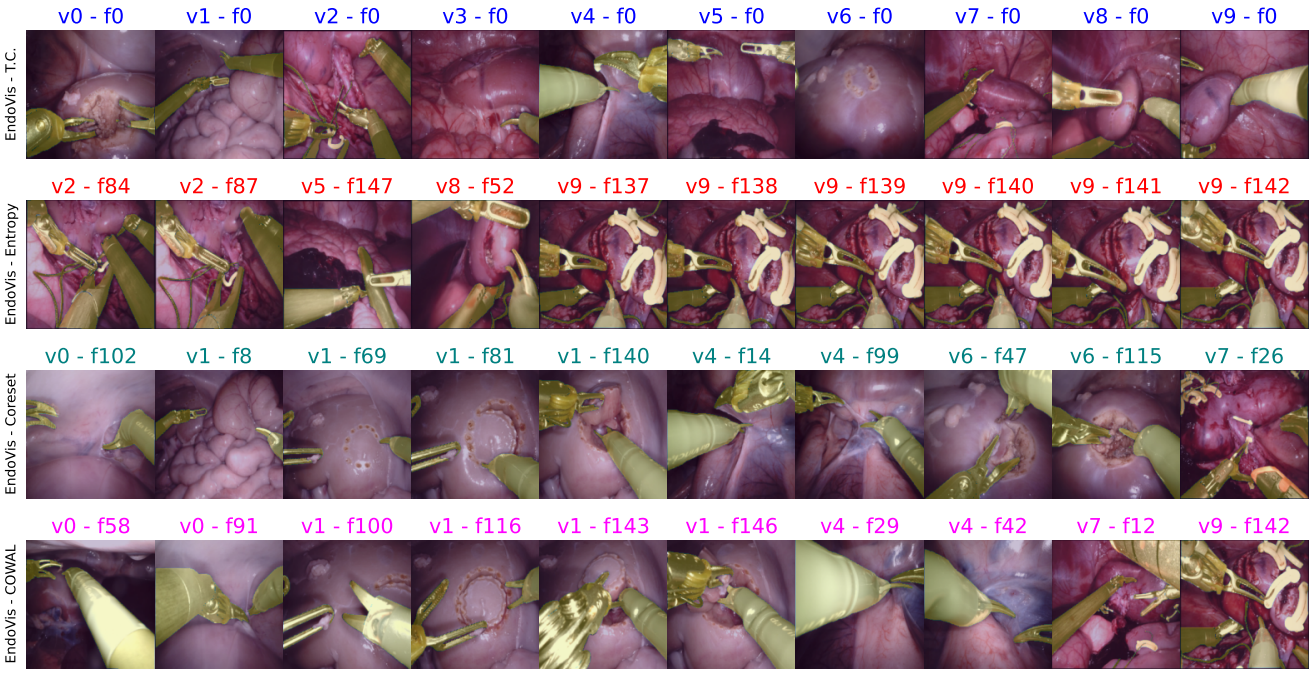

Figure 9. Selected frames at the first iteration on the EndoVis dataset. Video and frame numbers are indicated on top of each image.

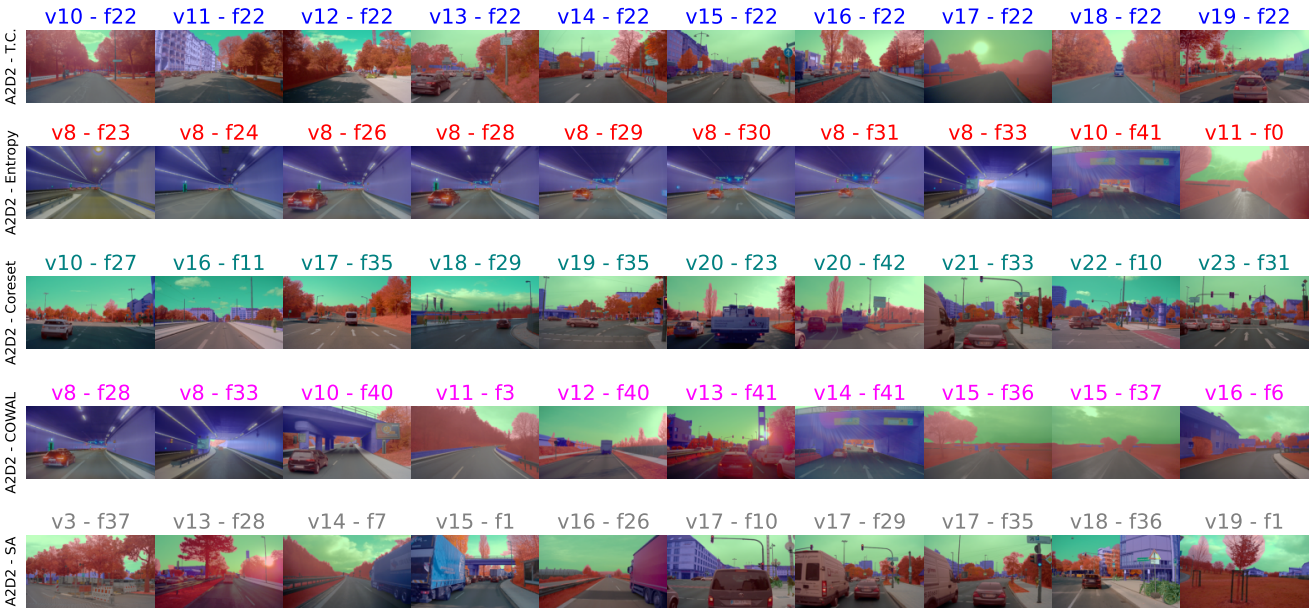

Figure 10. Selected frames at the first iteration on the A2D2 dataset. Video and frame numbers are indicated on top of each image.

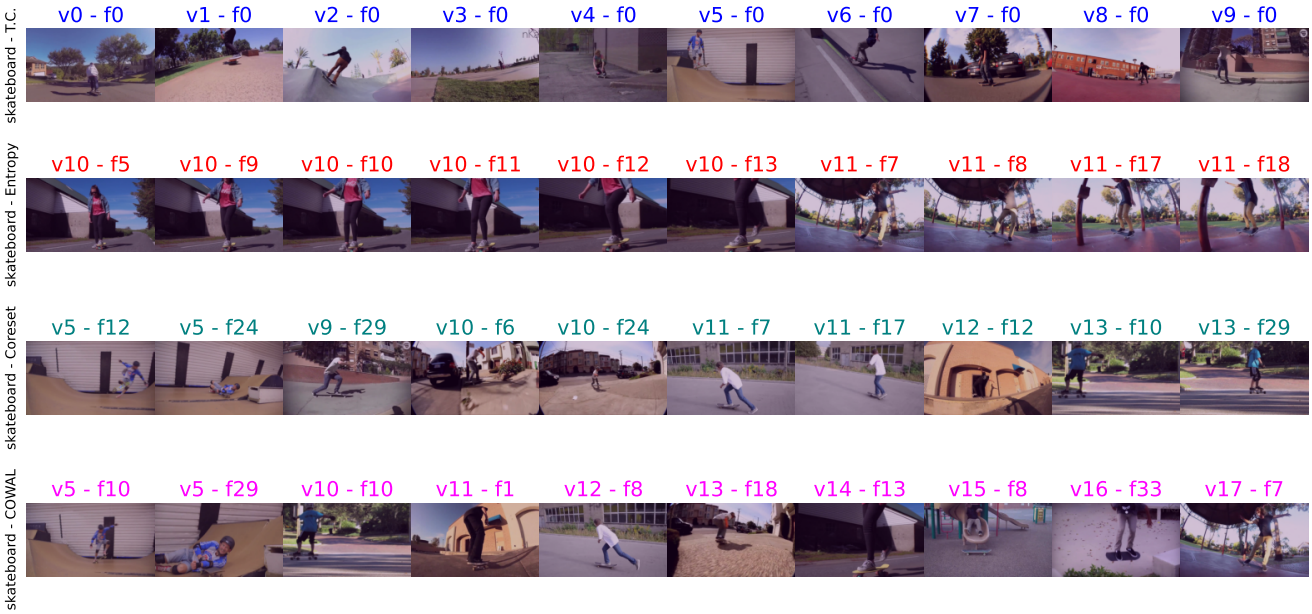

Figure 11. Selected frames at the first iteration on the Skateboard dataset. Video and frame numbers are indicated on top of each image.

Figure 12. Selected frames at the first iteration on the Parrot dataset. Video and frame numbers are indicated on top of each image.

References

[1] https://github.com/vainf/deeplabv3plus-pytorch. 6

[2] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Félix Fuentes-Hurtado, Evangello Flouty, Ahmed Kedir Mohammed, Marius Pedersen, Avinash Kori, Alex Varghese, Ganapathy Krishnamurthi, David Rauber, Robert Mendel, Christoph Palm, Sophia Bano, Guinther Saibro, Chi-Sheng Shih, Hsun-An Chiang, Juntang Zhuang, Junlin Yang, Vladimir Iglovikov, Anton Dobrenkii, Madhu Reddiboina, Anubhav Reddy, Xingtong Liu, Cong Gao, Mathias Unberath, Mahdi Azizian, Danail Stoyanov, Lena Maier-Hein, and Stefanie Speidel. 2018 robotic scene segmentation challenge. *CoRR*, abs/2001.11190, 2020. 1, 4, 5

[3] Max Allan, Alexey Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, Luis C. García-Peraza, Wenqi Li, Vladimir Iglovikov, Huoling Luo, Jian Yang, Danail Stoyanov, Lena Maier-Hein, Stefanie Speidel, and Mahdi Azizian. 2017 robotic instrument segmentation challenge. *CoRR*, abs/1902.06426, 2019. 1

[4] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *CoRR*, abs/1906.03671, 2019. 2

[5] Fan Bai, Xiaohan Xing, Yutian Shen, Han Ma, and Max Q.-H. Meng. Discrepancy-based active learning for weakly supervised bleeding segmentation in wireless capsule endoscopy images. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 24–34, Cham, 2022. Springer Nature Switzerland. 1

[6] Sima Behpour. Active learning in video tracking. *CoRR*, abs/1912.12557, 2019. 1, 2

[7] Javad Zolfaghari Bengar, Abel Gonzalez-Garcia, Gabriel Villalonga, Bogdan Raducanu, Hamed H. Aghdam, Mikhail Mozerov, Antonio M. López, and Joost van de Weijer. Temporal coherence for active learning in videos. *CoRR*, abs/1908.11757, 2019. 1, 2

[8] Sebastian Bodenstedt, Max Allan, Anthony Agustinos, Xiaofei Du, Luis Garcia-Peraza-Herrera, Hannes Kenngott, Thomas Kurmann, Beat Müller-Stich, Sebastien Ourselin, Daniil Pakhomov, Raphael Sznitman, Marvin Teichmann, Martin Thoma, Tom Vercauteren, Sandrine Voros, Martin Wagner, Pamela Wochner, Lena Maier-Hein, Danail Stoyanov, and Stefanie Speidel. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery, 2018. 1

[9] Samuel Budd, Emma C. Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *CoRR*, abs/1910.02923, 2019. 1, 2

[10] Lile Cai, Xun Xu, Jun Hao Liew, and Chuan Sheng Foo. Revisiting superpixels for active learning in semantic segmentation with realistic annotation costs. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10983–10992, 2021. 2

[11] Arantxa Casanova, Pedro O. Pinheiro, Negar Rostamzadeh, and Christopher J. Pal. Reinforced active learning for image segmentation. *CoRR*, abs/2002.06583, 2020. 2

[12] David M. Chan, Sudheendra Vijayanarasimhan, David A. Ross, and John F. Canny. Active learning for video description with cluster-regularized ensemble ranking. *CoRR*, abs/2007.13913, 2020. 1, 2

[13] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. 6

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. 4

[15] Pascal Colling, Lutz Roese-Koerner, Hanno Gottschalk, and Matthias Rottmann. Metabox+: A new region based active learning method for semantic segmentation using priority maps. *CoRR*, abs/2010.01884, 2020. 2

[16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016. 2

[17] Chengliang Dai, Shuo Wang, Yuanhan Mo, Elsa D. Angelini, Yike Guo, and Wenjia Bai. Suggestive annotation of brain tumour images with gradient-guided sampling. *CoRR*, abs/2006.14984, 2020. 1, 2

[18] Alireza Fathi, Maria-Florina Balcan, Xiaofeng Ren, and James M. Rehg. Combining self training and active learning for video segmentation. In *British Machine Vision Conference*, 2011. 1, 2

[19] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016. 2, 3, 5, 6, 7

[20] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. *CoRR*, abs/1703.02910, 2017. 2, 3, 5, 6, 7

[21] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: Audi Autonomous Driving Dataset. 2020. 2, 4, 5

[22] S. Alireza Golestaneh and Kris M. Kitani. Importance of self-consistency in active learning for semantic segmentation. *CoRR*, abs/2008.01860, 2020. 2

[23] Brent A. Griffin and Jason J. Corso. Bubblenets: Learning to select the guidance frame in video object segmentation by deep sorting frames. *CoRR*, abs/1903.11779, 2019. 1, 2

[24] David A. Gutman, Noel C. F. Codella, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Nabin K. Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin

imaging collaboration (ISIC). *CoRR*, abs/1605.01397, 2016. 7

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 4

[26] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning, 2011. 2

[27] Mobarakol Islam, Yueyuan Li, and Hongliang Ren. Learning Where to Look While Tracking Instruments in Robot-Assisted Surgery. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11768 LNCS, pages 412–420, 2019. 1

[28] Ziyi Jin, Tianyuan Gan, Peng Wang, Zuoming Fu, Chongan Zhang, Qinglai Yan, Xueyong Zheng, Xiao Liang, and Xuesong Ye. Deep learning for gastroscopic images: computer-aided techniques for clinicians. *BioMedical Engineering OnLine*, 21, 02 2022. 1

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. 6, 7

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 2

[31] Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *CoRR*, abs/1906.08158, 2019. 2, 3

[32] Gaston Lenczner, Adrien Chan-Hon-Tong, Bertrand Le Saux, Nicola Luminari, and Guy Le Besnerais. DIAL: deep interactive and active learning for semantic segmentation in remote sensing. *CoRR*, abs/2201.01047, 2022. 2

[33] Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. CEREALS - cost-effective region-based active learning for semantic segmentation. *CoRR*, abs/1810.09726, 2018. 2

[34] Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. Active learning by acquiring contrastive examples. *CoRR*, abs/2109.03764, 2021. 2, 3

[35] Sudhanshu Mittal, Joshua Niemeijer, Jörg P. Schäfer, and Thomas Brox. Best practices in active learning for semantic segmentation, 2023. 1, 2, 5

[36] Vishwesh Nath, Dong Yang, Bennett A. Landman, Daguang Xu, and Holger R. Roth. Diminishing uncertainty within the training pool: Active learning for medical image segmentation. *CoRR*, abs/2101.02323, 2021. 1

[37] Vishwesh Nath, Dong Yang, Holger R. Roth, and Daguang Xu. Warm start active learning with proxy labels & selection via semi-supervised fine-tuning, 2022. 1

[38] Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Reza Haffari, Anton van den Hengel, and Javen Qinfeng Shi. Active learning by feature mixing, 2022. 2, 3

[39] Haonan Peng, Shan Lin, Daniel King, Yun-Hsuan Su, Randall A. Bly, Kris S. Moe, and Blake Hannaford. Reducing annotating load: Active learning with synthetic images in surgical instrument segmentation. *CoRR*, abs/2108.03534, 2021. 1, 2

[40] Aneesh Rangnekar, Christopher Kanan, and Matthew Hoffman. Semantic segmentation with active semi-supervised learning, 2022. 2

[41] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach, 2017. 2, 3, 5, 6, 7

[42] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948. 2, 3, 5, 6, 7

[43] X. Shi, Y. Jin, Q. Dou, and P. Heng. Lrtd: long-range temporal dependency based active learning for surgical workflow recognition. *Int J Comput Assist Radiol Surg*, 15(9):1573–1584, 2020. 1

[44] Xueying Shi, Yueming Jin, Qi Dou, and Pheng-Ann Heng. LRTD: long-range temporal dependency based active learning for surgical workflow recognition. *CoRR*, abs/2004.09845, 2020. 2

[45] Yejee Shin, Taejoon Eo, Hyeongseop Rha, Dong Jun Oh, Geonhui Son, Jiwoong An, You Jin Kim, Dosik Hwang, and Yun Jeong Lim. Digestive organ recognition in video capsule endoscopy based on temporal segmentation network. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 136–146, Cham, 2022. Springer Nature Switzerland. 1

[46] Alexey Shvets, Alexander Rakhlin, Alexandr A. Kalinin, and Vladimir Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. *CoRR*, abs/1803.01207, 2018. 1

[47] Yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. *CoRR*, abs/1911.11789, 2019. 1, 2, 6

[48] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. *CoRR*, abs/1904.00370, 2019. 1, 2, 3, 6

[49] Asim Smailagic, Hae Young Noh, Pedro Costa, Devesh Walawalkar, Kartik Khandelwal, Mostafa Mirshekari, Jonathon Fagert, Adrian Galdran, and Susu Xu. Medal: Deep active learning sampling method for medical image analysis. 09 2018. 1

[50] Deepthi Sreenivasaiah and Thomas Wollmann. MEAL: manifold embedding-based active learning. *CoRR*, abs/2106.11858, 2021. 2

[51] Alaa Tharwat and Wolfram Schenck. A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics*, 11(4), 2023. 2

[52] Carl Vondrick and Deva Ramanan. Video annotation and tracking with active learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. 1, 2

[53] Tsung-Han Wu, Yueh-Cheng Liu, Yu-Kai Huang, Hsin-Ying Lee, Hung-Ting Su, Ping-Chia Huang, and Winston H. Hsu. Redal: Region-based and diversity-aware active learning for point cloud semantic segmentation. *CoRR*, abs/2107.11769, 2021. 2

[54] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. *CoRR*, abs/2111.12940, 2021. 1, 2

[55] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas S. Huang. Youtube-vos: A large-scale video object segmentation benchmark. *CoRR*, abs/1809.03327, 2018. 4, 5

[56] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z. Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. *CoRR*, abs/1706.04737, 2017. 1, 2, 3, 5, 6, 7

[57] LeCun Yann and Cortes Corinna. The mnist database of handwritten digits, 1998. 7

[58] Ghada Zamzmi, Tochi Oguguo, Sivaramakrishnan Rajaraman, and Sameer Antani. Open world active learning for echocardiography view classification. volume 12033, page 20, 04 2022. 1

[59] Ziyuan Zhao, Zeng Zeng, Kaixin Xu, Cen Chen, and Cuntai Guan. Dsal: Deeply supervised active learning from strong and weak labelers for biomedical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25(10):3744–3751, 2021. 1, 2