VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Improving Machine Learning Performance using Ensemble Methods with Quality Assurance

Doctoral Thesis Proposal

Mgr. David Číž

Supervisor: doc. Mgr. Vít Vondrák, Ph.D.

Ostrava, August 2024

Contents

Chapter 1

Introduction

Image segmentation stands as a cornerstone process in computer vision, involving the intricate task of partitioning an image into multiple segments or regions, each corresponding to a meaningful part of the image. This critical step in image analysis forms the bedrock for a multitude of applications spanning diverse domains, from medical imaging and autonomous driving to satellite imagery analysis and industrial quality control.

The ability to accurately segment images carries profound implications across various fields. In the realm of medical imaging, precise segmentation can significantly aid in tumor detection, organ measurements, and treatment planning, potentially revolutionizing patient care. For autonomous vehicles, segmentation proves crucial in understanding the environment, detecting obstacles, and ensuring safe navigation. Satellite imagery benefits from segmentation in mapping, urban planning, and monitoring environmental changes, contributing to more informed decision-making in these areas. Industrial applications harness segmentation for defect detection and quality assurance in manufacturing processes, enhancing efficiency and product quality.

As the volume and complexity of visual data continue to grow at an unprecedented rate, the demand for robust, efficient, and accurate segmentation techniques becomes increasingly critical. This exponential growth in data presents both challenges and opportunities for researchers in the field. Despite significant advancements in recent years, particularly with the advent of deep learning techniques, image segmentation continues to grapple with several challenges. Real-world scenarios present complexities such as occlusions, varying scales, and diverse object orientations that push the boundaries of current segmentation methods. Striking a balance between accuracy and real-time performance remains a persistent challenge, especially in applications demanding swift processing. If machine learning-based segmentation methods shall be considered, the development of models capable of performing well across different domains and datasets poses another significant hurdle. Moreover, the creation of large-scale, high-quality annotated datasets for training purposes presents logistical and resource-related challenges. Ensuring consistent performance under varying lighting conditions and image qualities further complicates the landscape of image segmentation research.

This thesis aims to address some of these challenges by exploring novel approaches to utilizing deep learning architectures for image segmentation, with a particular focus on improving both accuracy and reliability. While ensemble methods have been successfully applied in image segmentation, this research will investigate innovative ways to combine ensemble techniques with quality assessment methods, particularly to obtain segmentation data of high quality (accuracy) by combining several existing segmentation results. That way, for example, computer-generated annotation data of quality close to that of human-created ones could be created. But, unlike that from humans, the approach could deliver significantly larger numbers of annotations which could be useful for learning-based segmentation methods. While this is a general idea that could be utilized in various domains, we plan to develop and demonstrate it in the field of time-lapse biomedical images that are showing multitudes of cells or nuclei acquired using optical microscopy techniques.

The subsequent chapters of this thesis will provide a comprehensive exploration of the state of the art in image segmentation, detail the proposed methodologies, present experimental results, and discuss the implications and future directions of this work. Through this research, we aim to push the boundaries of what is possible in image segmentation, contributing to the broader field of computer vision and its myriad applications.

Chapter 2

Machine Learning in Image Segmentation

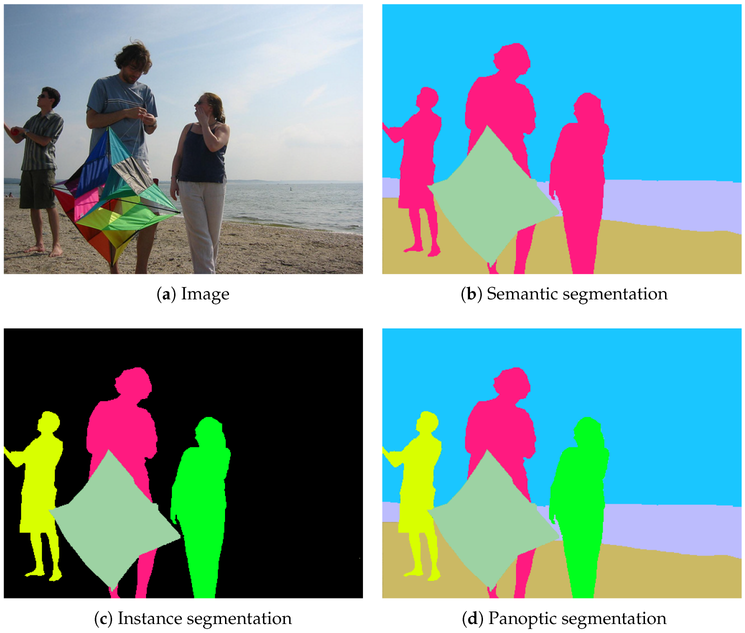
2.1 Image Segmentation

Image segmentation is a fundamental process in computer vision that involves partitioning an image into multiple segments or regions, each corresponding to a meaningful part of the image. For example, outlining all cars in the digital photograph of a street is segmentation of cars, or identifying pixels that make up a picture of cell nucleus in a digital micrograph is cell nucleus segmentation. The result of this process is typically a segmentation mask or multiple masks. The goal is to simplify the representation of an image to make it more meaningful and easier to analyze [1]. The field has seen significant advancements in recent years, particularly with the advent of deep learning techniques. There are multiple ways of categorizing segmentation approaches, but here we distinguish them based on their end goal, see also figure 2.1.

Semantic Segmentation Semantic segmentation aims to categorize every pixel in an image and assign to each a value that is essentially an ID of a semantic class to which the pixel (and the object) belongs. The IDs are often termed class labels. This approach, however, doesn't differentiate between individual instances of the same class.

Instance Segmentation Instance segmentation goes a step further by differentiating between multiple objects of the same class, assigning different values to each of them. Unlike semantic segmentation, instance segmentation doesn't necessarily assign a label to every pixel in the image.

Panoptic Segmentation Panoptic segmentation combines the approaches of semantic and instance segmentation. In this method, each pixel belongs to a class of objects, and the objects themselves are differentiated as distinct instances. This provides a comprehensive understanding of the scene.



(**a**) Image

(**b**) Semantic segmentation

(**c**) Instance segmentation

(**d**) Panoptic segmentation

Figure 2.1: Visual representation of different types of image segmentation: (a) Input image, (b) Semantic segmentation, (c) Instance segmentation, and (d) Panoptic segmentation. Adapted from [2].

2.2 Segmentation Techniques

The field of image segmentation has been revolutionized by deep learning techniques in recent years. While traditional methods such as thresholding, region-based segmentation, edge detection, and contouring were once common.

The predominant approaches now utilize deep Neural Networks (NNs)[3]. Fully Convolutional Networks (FCN) pioneered the use of end-to-end trained Convolutional Neural Networks (CNNs) for semantic segmentation, allowing for arbitrary-sized inputs [4]. Mask R-CNN extended the Faster R-CNN object detection framework to include instance segmentation, effectively combining object detection and segmentation [5]. DeepLab introduced atrous convolutions for semantic segmentation, allowing for larger receptive fields without increasing computational cost [6]. U-Net, originally developed for biomedical image segmentation, features a symmetric encoder-decoder structure with skip connections, which has found wide applicability beyond its original domain [7].

More recent developments include the YOLO (You Only Look Once) family of models, which, while initially designed for object detection, now incorporate segmentation capabilities, offering real-time performance [8]. The Segment Anything Model (SAM), developed by Meta AI, represents a significant leap forward, capable of segmenting any object in an image based on various types of prompts [9].

These deep learning approaches have significantly improved the accuracy and efficiency of image segmentation tasks across various domains, including medical imaging [10], autonomous driving [11], and satellite imagery analysis [12].

2.3 Challenges in Segmentation

Despite recent advancements, image segmentation still faces several challenges:

- Dealing with complex scenes and occlusions,
- Handling varying scales and orientations of objects,
- Real-time performance for online collected data to extract just-in-time meaningful information,
- Addressing the need for large annotated datasets for training,
- Ensuring robustness to different lighting conditions and image qualities,
- Developing generic segmentation models capable of handling diverse object types, reducing dependency on specialized datasets,
- Evaluating performance without ground truth data,
- Effectively combining multiple segmentation models to leverage their individual strengths and improve overall accuracy.

Addressing these challenges is an active area of research in the field of computer vision and machine learning. In this work, I focus on exploring the quality assessment of the computed masks and their influence on the quality of ensemble methods, with the goal of improving the final segmentation masks.

Chapter 3

Quality Assessment and Ensembles

3.1 Quality Assessment of Segmentation

Quality assessment is a process with the goal of determining if the studied system has the desired capabilities and to what extent. In the realm of segmentations, the goal is to determine the quality, robustness and reliability of the given model [13]. The terms quality assessment, quality assurance (QA), and quality control are often used in similar contexts throughout the literature. For the purposes of this work, while I recognize the distinctions between these terms, I use them in a closely related manner, interchangably. Specifically, I employ quality assessment as a key component of a broader quality assurance process. By evaluating the quality of individual segmentations (quality assessment), and incorporaring this information during the ensembling process, I aim to improve the overall quality of the final segmentation results (quality assurance). This approach allows us to perform a form of quality control, selecting and fusing only the high-quality segmentations to produce a more reliable final output. Thus, when I refer to QA in this work, I'm often describing a process that begins with quality assessment and culminates in a form of quality control, all serving the overarching goal of ensuring and obtaining segmentation results of better quality.

3.1.1 Quality Metrics and Ground Truth Data

To evaluate the quality of the segmentation masks, multiple metrics and approaches are used [14]. Provided we have access to the ground truth data, the following metrics are usually used, giving us a concrete measurement. The Intersection over Union (IoU) and Dice coefficient are two of the most prevalent metrics to measure the accuracy of the shape of segmentation masks. The IoU, also known as the Jaccard index, measures the overlap between the predicted segmentation and the ground truth. The Dice coefficient, which is closely related to IoU and works similarly, is particularly useful in medical image analysis due to its sensitivity to small variations in segmentation. Other important metrics, that also work at the pixel accuracy, are based on calculating the proportion of correctly and incorrectly classified pixels, which is typically further aggregated into recall, specificity, and precision. Another such example is the popular F1 score, which provides a balanced measure of precision and recall. It's crucial to note that class imbalance in datasets can significantly affect these metrics, potentially leading to misleading interpretations of model performance.

3.1.2 Metrics with no Ground Truth Data

In the case of a segmentation on data without ground truth, different approaches are required. One approach is to judge the segmentation final shape based on known characteristics of the domain (Contrast, Dissimilarity, Homogeneity) using evaluation metrics. While unsupervised approaches are used [15], recent developments use semi-supervised or supervised approaches. Recent years have seen a shift towards deep learning models capable of self-evaluation, often through the incorporation of confidence and uncertainty maps. These maps, typically derived from the final layer of neural networks, provide a visual representation of the model's certainty in its segmentations and are used to enhance quality assessment evaluations. Notable example is [16] where they obtained uncertainty features from a Bayesian Swin transformer-based U-Net uncertainty maps to train a random forest classifier for quality evaluation. Similarly, [17] and [18] used probability and uncertainty maps as inputs to their QA network. The authors in [19] introduced a quality control algorithm based on fuzzy uncertainty, combining test-time augmentation and Monte Carlo dropout to quantify segmentation quality. The caveat of this approach is that the model is evaluating itself, therefore very confident model yields many false positives and negatives. Another approach in evaluating the performance of a model in lieu of ground truth is the use of a seperate model, that is trained to recognize the quality of the segmentation given the source image and the segmentation. This is circumvented in [20] where their proposed QA-Net based Rib-Cage architecture does regression, using only the source image and segmentation mask. Similar approach can be seen in [21], where they also used this input combination trained U-net based CNN for detection and localization of errors in medical images, facilitating faster expert reevaluation. For training these QA networks, the training data needs to contain both examples of good and bad segmentations. For this purpose [22] used early epoch segmentations as examples of poor segmentation when training their quality control regression model.

Expanding on these ideas, [23] developed a two-step process of reconstruction followed by assessment of the difference with the original image. In an effort to mitigate false negatives in this approach, [24] attempted to suppress their effect by using a structural similarity index measure based loss function. Building on these reconstruction-based approaches, [25] introduced a method that generates self-reflective reference images from segmentation masks, coupled with a difference investigator equipped with a semantic class-aware compactness constraint. Leveraging recent advancements in foundation models, [26] created prompts for the Segment Anything Model (SAM) from extracted segmentations and compared the results to their original segmentation to derive a quality score. This innovative use of large-scale pre-trained models opens up new possibilities for quality assessment. Other researchers have explored techniques that rely on reverse engineering and model adaptation. In [27] proposed Reverse Classification Accuracy, training a segmentation model on a single new image without ground truth, then using that model to segment a set of known images with ground truths in a database. The premise is that if the database contains something similar to the new image, it should provide a reasonably accurate segmentation, which can then be used to assess the quality of the original segmentation. Building on this concept, [28] developed a variation of this approach for Real-Time predictions of cardiovascular MR segmentations.

Adjacent fields, such as contour QA, have also contributed valuable insights. To name a few, [29] developed a focused, QA framework for contours, addressing a problem analogous to segmentation quality assessment. Similarly, [30] introduced custom metrics derived from geometric attribute distributions, specifically tailored for radiology data. In [31] they trained decision trees based on texture constraints. These approaches demonstrate the potential of leveraging domain-specific knowledge to enhance QA processes.

As segmentation tasks become increasingly complex, particularly in specialized fields, these innovative QA techniques play a crucial role in ensuring the reliability and accuracy of segmentation results.

3.2 Ensemble Techniques for Machine Learning

Ensemble methods combine multiple models to create a more robust and accurate predictive system. These techniques have shown remarkable success across various domains, including medical image analysis [32], natural language processing [33], and bioinformatics [34]. In the context of image segmentation, ensembles can help overcome limitations of individual models, leading to more precise and reliable segmentation results as explored in [35]. Several approaches to model ensembling exists, each with its unique characteristics and applications:

Bagging, or also Bootstrap Aggregating, involves training multiple models on different subsets of the training data, typically drawn with replacement. The outputs of these models are then combined through voting or averaging. Bagging helps reduce overfitting and variance in the predictions. Popular bagging algorithm is Random Forest.

Boosting, means weak learners are trained iteratively, with each subsequent model focusing on the errors of the previous ones. Popular boosting algorithms include AdaBoost and Gradient Boosting Machines.

Stacking, also known as stacked generalization, this method uses the outputs of multiple models as inputs to a meta-model that learns how to best combine them. Stacking can capture complex relationships between base models and often outperforms simpler ensemble methods.

Bucket of models involves creating a diverse set of models and selecting the most appropriate one for each input or task, and training that model on the input.

3.2.1 Dynamic Ensembles

Dynamic ensembles represent an advanced approach to combining models, where the ensemble composition or weighting is adjusted based on the input or context. This adaptive nature allows dynamic ensembles to leverage the strengths of different models in varying scenarios, potentially leading to improved performance across a wide range of inputs [36, 37]. The following are key principles and requirements of dynamic ensembles:

Base Learner A base learner, also known as a base model or weak learner, is an individual machine learning model that forms part of the ensemble. These can be of the same type (e.g., all decision trees) or different types (e.g., a mix of decision trees, neural networks, and support vector machines).

Meta Learner A meta learner, or combiner, is a higher-level model that learns how to optimally combine the predictions of the base learners. In dynamic ensembles, the meta learner often plays a crucial role in determining which base learners to use or how to weight their outputs for each input.

Dynamic Selection It is a process of choosing a subset of base learners from the available pool for each new input or set of inputs. The selection can be based on various criteria such as the estimated competence of each base learner for the given input, diversity of the selected models, or historical performance on similar inputs.

Competence Estimation Estimating the competence of each base learner for a given input can be done using techniques like k-Nearest Neighbors (k-NN) based approaches, Meta-learning techniques or Oracle-based methods.

Diversity Preservation Maintaining diversity among the selected base learners is important to ensure the ensemble captures a wide range of perspectives on the input data.

Adaptivity The ensemble should be able to adapt to changes in the input distribution or concept drift in streaming data scenarios.

Efficiency The selection or weighting process should be computationally efficient to allow for real-time or near real-time application.

Ensemble pruning It is a process that aims to improve the performance and efficiency of an ensemble by only selecting a subset of base learners.

3.2.2 Ensemble Strategies

Many methods have been reported in literature for ensembling strategies and how to best combine the models outputs. As a simple example, [38] used averaging to achieve state of the art. For polyp segmentation in colonoscopy images, a simple bitwise combination has also shown promise [39]. Majority voting algorithm using posterior probablities from the last layer or neural networks are used in [40] for tumor segmentation. A more advanced algorithm for ensembling skin lesions are used in [41], where a dual-threshold method combines mean probabilities across CNNs with blob analysis. Weighted segmentation ensembles with weights determined by swarm algorithms have emerged as another effective technique [42]. Recent research explored weighted average ensembling and Bayesian voting, combining models probability maps [43].

Instance segmentations ensembles first need to find out which instances between different models match the same object. For example in [44], they propose SISE (Simple Instance Segmentation Ensemble) algorithm, that combines objects that have among themselves the highest IoU by means of disjunction (logical or). In another approach, 2D mask slice fusing strategy using confidence scores has been explored in [45] where they also utilize ensemble learning for training without any ground truth. Sequential boosting approaches have also been used with confidence scores to train subsequent model generations, iteratively improving performance [46].

Dynamic ensembles have gained traction, with applications in imbalanced datasets, data streams, and regression tasks [47, 48]. Strategies for ensemble pruning with meta-learners for classification tasks based on multiple criteria (Neighbors' hard classification, Posterior probability, Overall local accuracy, Output profile classification, Classifier's confidence) are used in [49]. In [50], they used a multi-label classification process for the selection of base learners. Information entropy-based optimization approaches to ensemble pruning have been developed, offering a more general framework [51]. Frameworks also exist for easy access to common dynamic ensemble selection techniques: DESReg [52] for regression tasks and DESlib [53] for classification tasks.

3.3 Challenges in Quality Assurance and Ensembling

Recent advancements in image segmentation have increasingly focused on combining ensemble methods with QA techniques. Several notable studies have paved the way for our research. Previous work [54, 55] introduced a voting scheme to combine segmentation masks, coupled with a QA step that predicts Dice scores to select the optimal mask. Other researchers [56, 57] explored dynamic ensemble selection in medical imaging, particularly for glaucoma detection. Their approach not only applied ensemble methods to segmentation but also extracted metrics from the segmentations for further ensemble processing. A comprehensive framework [58] incorporating both ensembles and quality control has been presented. This method selects the final segmentation from multiple candidates and several on-the-fly created ensembles based on a quality score.

While significant progress has been made, there remains an unexplored area at the intersection of these approaches, particularly in the context of dynamic ensemble selection for segmentation tasks. Current ensemble methods typically use a fixed set of inputs from different models and processing techniques, without preliminary screening for low-quality masks. Although the weighted averaging in these ensembles may mitigate the impact of poor-quality inputs to some extent, I propose a more proactive approach: a pre-filtering step that dynamically determines the candidates for the ensemble for each new input by ensemble pruning techniques. This ensemble pruning will be done by a segmentation QA methods. Ensemble prunings for segmentation problems is very underdeveloped area of research, it's usually only applied to classification tasks.

Chapter 4

Novel Approach to Quality Assurance and Ensembling

4.1 Empirical Observations

A motivation to consider QA has emanated from our recent research medical domain project in which we constructed machine learning pipelines for classification and segmentation tasks. We have observed a considerable label noise in our datasets. For example, 2.17 % of images were duplicates, which could have been revealed only by calculating image hash maps. This is something that good QA could have quickly identified for us. Additionally, there was a huge disbalance ratio in classes, with 98.64 % being background. Several models were created, but due to the misbalance special care had to be made to ensure proper sensitivity and precision scores, as the IoU score for models returning only background was 0.55, which is usually considered a good score. This metric value was not very different from better-performing models. Dynamic ensembling of these models could also be improved by additional quality checks during inference, which are especially important in a clinical setting. Additional and better QA methods would also identify these labeling problems sooner, and this topic is in the project recently being improved with tools like [59]. In another research project, in which we were tracking zebrafish animal, we were employing a regression network to return head and tail coordinates in the image. During labeling, the two coordinates were sometimes swapped, leading to incorrect labelings, which again led to the need for better automatic checks of labels. During sudden erratic movements near the edges of the image, the tracking would sometimes fail, moving it very far from the prediction of the previous frame. This leads us to consider a more monitoring role for QA, which should also include temporal consistency and shape checks for these cases.

I was also involved in another research, where I segmented carbon dots into semantic segmentation, before the additional step of transforming it into instance segmentation using StarDist [10]. During the semantic segmentation phase, we encountered challenges due to the diverse background, which often included a mesh with varying shapes and color intensities. This variability occasionally led to inaccurate segmentations. Good quality metrics need to be used to automatically detect these cases, based on the general shape and size of carbon dots in the image.

Additionally, through collaboration, I have been granted access to both train and test data from the 2020 Cell Tracking Challenge, along with all competitors' submissions and evaluations. This access aims to explore additional approaches for creating improved silver-quality reference segmentation annotations. The cell tracking challenge spans a wide range of cell types and imaging modalities, providing a comprehensive benchmark for cell segmentation and tracking algorithms. These datasets align well with my proposed research objectives, as they contain both semantic and instance segmentations, temporal information in the form of cell segmentations, and require ensembling of several models' results.

4.2 Cell Tracking Challenge Data

The data comprises 16 diverse datasets, each containing contributions from multiple competitors. These contributions are filtered based on quality assessment across the dataset. The following is transcribed from information in [60] for additional context.

The cell tracking challenge employs a sophisticated two-stage **fusion algorithm** to generate silver standard segmentation annotations. In the first stage, models are evaluated based on their average Jaccard score across all images in a test dataset. Those falling below a certain threshold are removed from the pool of solutions for that dataset. This filtering is applied at the model level, with no additional screening of individual segmentations. In the second stage, the remaining models that pass the quality threshold are used in a fusion algorithm, a modified version of the label fusion approach [38]. This algorithm begins by selecting up to 16 best-performing methods based on segmentation and detection scores above 50% of human performance on both training and test datasets. For each marker in the detection gold truth, segments from selected results that overlap the majority of the marker are collected. These collected segments are then fused using pixel-wise voting, where pixels that appear in more than a threshold of the segments are retained. This threshold is optimized for each video to maximize segmentation scores relative to the gold standard. In cases where fewer segments are available than the threshold, the segment from the best-performing method is used. To resolve overlaps, a marker-based watershed method finds boundaries and creates touching segments from overlapping ones. Final adjustments include replacing segments with high overlap (>20%) with corresponding detection markers.

This fusion algorithm aims to combine the strengths of multiple high-performing models to produce a more accurate segmentation, as can be seen in table 4.1, the fusing ensemble improves the resulting segmentation above the best individual score. However, there are also limitations to this approach. The initial filtering is based on overall model performance, potentially discarding models that perform well on specific subsets of data. Furthermore, the approach does not consider the quality of individual segmentations, which may vary within a single model's output.

The proposed research aims to address these limitations by introducing a more granular quality assessment approach and a dynamic ensemble selection method. Further analysis and proposed architecture of the QA architecture is further discussed in 4.3. Same for our ensembling approach in 4.4.

4.3 Quality Assurance Network

The first hypothesis posits that the current fusion method can be improved by filtering out low-quality object instances. To test this, a human-guided selection process was implemented to identify high-quality segmentation inputs. These curated inputs were then fed into the fusion method and compared against the existing solution. The results of this comparison are presented in Table 4.1. In most cases, it was found that this selective approach significantly improved upon the CTC Baseline, validating the hypothesis. Building on these findings, the next step is to design a QA network capable of performing this filtering process automatically.

As discussed above, numerous QA approaches exist for segmentation. Given the constraint of only having access to binary results rather than the models themselves or uncertainty maps, it is necessary to adapt existing solutions to this specific problem. Approaches like GAN-based methods or Reverse Classification Accuracy could potentially be modified for this use case. Notably, QANet [20] was trained on data from the cell tracking challenge, making it particularly relevant to this domain.

Training these networks on instance segmentation evaluation poses certain challenges, as most existing methods are designed for semantic segmentation and do not consider true and false negatives. Obtaining poor segmentations for training to recognize them is addressed in some studies, but the network also needs to train on inputs with incorrect proposed segmentations where there are no cells. As we lack ground truth for every cell in a given image, special care must be taken to provide examples that truly do not correspond to an existing cell. Fortunately, tracking information can be utilized for this purpose, indicating areas where there are cells without ground truth segmentations.

Drawing from my previous experience in tracking **[61, 62]**, I propose to investigate a less explored area: the utilization of temporal information in building and training quality assessment methods for segmentation tasks. I hypothesize that sudden changes in otherwise continuous segmentations of image series may signal potential problems more effectively than static assessments. These temporally inconsistent masks may provide worse overall performance compared to masks from models with coherent, stable segmentations over time.

To leverage this temporal information, I propose implementing a Temporal Consistency Scoring system within the QA network. This system would analyze instance segmentations across a sliding

Dataset_Seq	SIMPLE	STAPLE	CTC Baseline	Best Ind.	Ch.P. CTC	Segments
BF-C2DL-HSC_01	0.898	0.895	0.898	0.895	0.902	80
BF-C2DL-HSC_02	0.853	0.853	0.853	0.847	0.886	48
BF-C2DL-MuSC_01	0.820	0.823	0.833	0.814	0.871	39
BF-C2DL-MuSC_02	0.784	0.784	0.784	0.776	0.836	80
DIC-C2DH-HeLa_01	0.961	0.961	0.965	0.961	0.966	98
DIC-C2DH-HeLa_02	0.961	0.961	0.965	0.952	0.966	97
Fluo-C2DL-MSC_01	0.779	0.766	0.792	0.729	0.796	112
Fluo-C2DL-MSC_02	0.762	0.746	0.817	0.762	0.837	116
Fluo-C3DH-A549_01	0.975	0.991	0.991	0.991	-	-
Fluo-C3DH-A549_02	0.954	0.991	0.991	0.991	-	-
Fluo-C3DL-MDA231_01	0.729	0.718	0.739	0.684	0.661	120
Fluo-C3DL-MDA231_02	0.734	0.738	0.752	0.705	0.738	103
Fluo-N2DH-GOWT1_01	0.913	0.920	0.933	0.913	0.941	144
Fluo-N2DH-GOWT1_02	0.976	0.976	0.976	0.954	0.977	121
Fluo-N2DL-HeLa_01	0.874	0.872	0.884	0.860	0.904	491
Fluo-N2DL-HeLa_02	0.901	0.898	0.909	0.891	0.925	895
Fluo-N3DH-CHO_01	0.867	0.853	0.883	0.852	0.927	115
Fluo-N3DH-CHO_02	0.933	0.928	0.931	0.920	0.957	117
Fluo-N3DH-CE_01	N/A	N/A	0.705	0.745	0.785	101
Fluo-N3DH-CE_02	N/A	N/A	0.711	0.725	0.771	100
Fluo-C3DH-H157_01	0.845	0.891	0.901	0.891	0.891	119
Fluo-C3DH-H157_02	0.799	0.779	0.824	0.779	0.879	99
PhC-C2DL-PSC_01	0.779	0.779	0.785	0.764	0.818	242
PhC-C2DL-PSC_02	0.780	0.780	0.784	0.779	0.819	254
PhC-C2DH-U373_01	0.949	0.948	0.951	0.947	0.953	100
PhC-C2DH-U373_02	0.884	0.882	0.888	0.880	0.897	110

Table 4.1: Comparison of cherry-picked inputs (Ch.P. CTC) to different fusing algorithms. Cherry picking was conducted on the denoted number of segments.

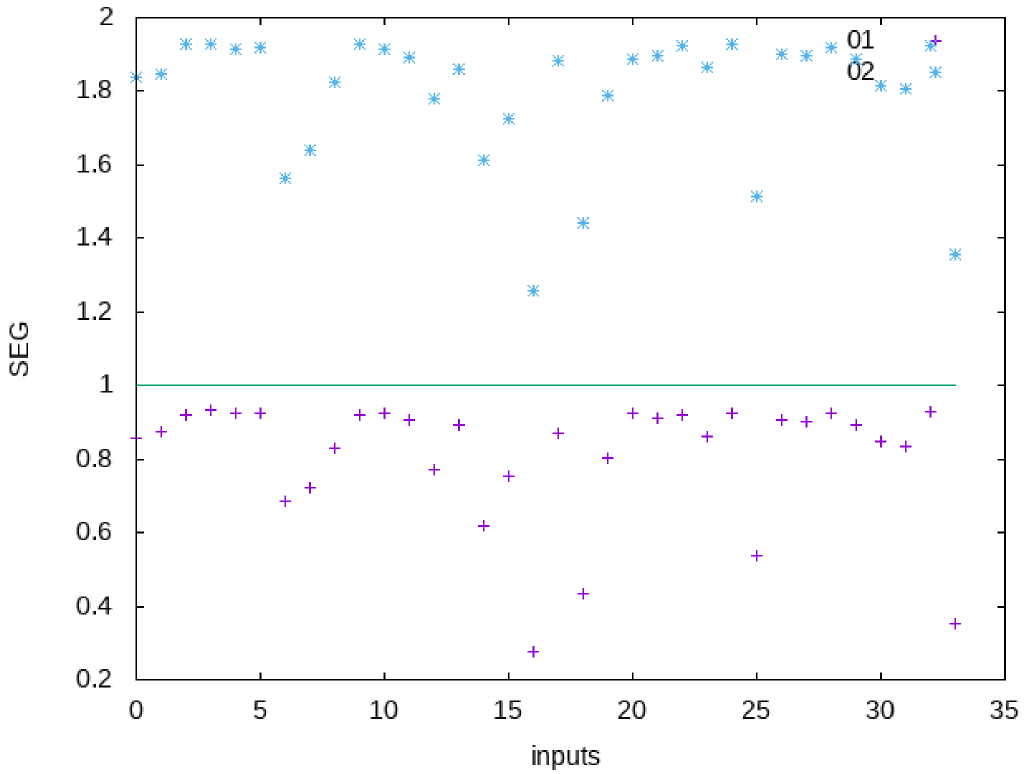


Figure 4.1: IoU scores across frames for PhC-C2DH-U373 dataset. Lower points ('+') show sequence 01, upper points ('*') show sequence 02 shifted up by 1. X-axis: frames; Y-axis: IoU (0-1 for seq. 01, 1-2 for seq. 02).

time window, calculate IoU score for each instance between consecutive frames, and compute a temporal consistency score based on the average IoU over the time window.

Using these temporal consistency scores, the QA network can filter out instances with low temporal consistency, potentially indicating segmentation errors. The system can adapt its quality thresholds based on the temporal context. For instance, it might apply stricter criteria during periods of stability and more lenient ones during expected change events (like cell division). A weighted ensemble approach can be implemented, giving more importance to instances with higher temporal consistency. Additionally, the system can provide feedback to individual segmentation models, flagging those that consistently produce low-quality instances for potential retraining.

By incorporating this temporal aspect into the QA process, This approach is particularly well-suited to the cell tracking application, where the temporal continuity of cell instances is a key feature of the data. An analysis of how IoU changes over time has been conducted with results visible in Figure 4.1. Both sequences show fluctuations in IoU scores across frames with sudden drops. This effect can be seen across all datasets.

The QA architecture approach can be seen in Figure 4.2. This architecture will be adaptable to

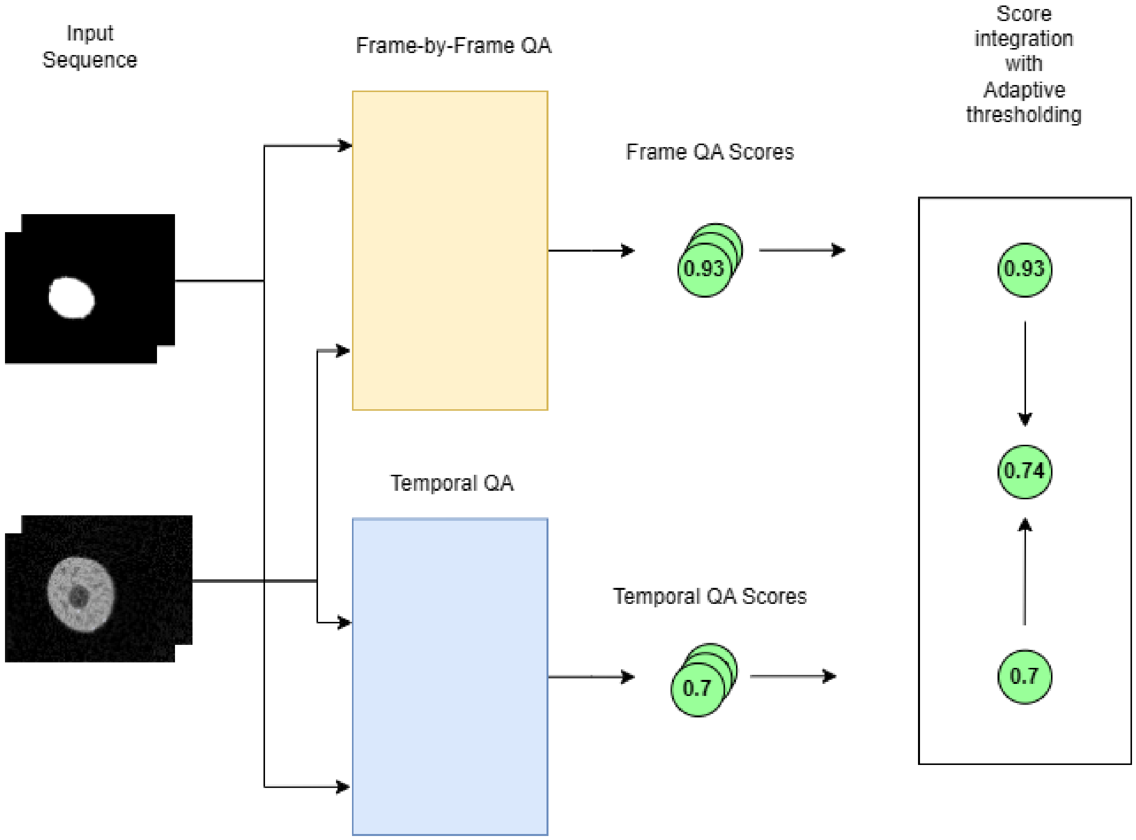


Figure 4.2: QA architecture with Frame-by-Frame QA, Temporal QA and Adaptive thresholding. Input combination of segmentations and source images are scored independently by both QA methods and then weighted and adaptively thresholded.

incorporate and test different approaches, such as passing both QA scores seperately to the ensemble as additional context.

4.4 Dynamic Ensemble Architecture for Variable Input Size

The challenge of incorporating a variable number of inputs in ensemble or fusion methods for image segmentation presents several potential approaches. One viable strategy involves designing a model with a maximum number of inputs (n_max), where not all inputs are necessarily utilized in every instance. This approach can be implemented using a specialized ignore mask, allowing the model to disregard certain inputs as needed. To ensure input order invariance, weight or parameter sharing can be employed across all inputs a method often used in siamese neural networks [63]. Additionally, a fixed source image input could be incorporated into this design, resulting in a modified version of existing architectures.

An alternative approach could involve the use of recurrent segmentation networks, such as LSTM-UNets [64], which have demonstrated success in various challenges. However, this method presents challenges related to the inherent order-dependence of RNNs, potentially leading to information loss or diminished influence of earlier segmentations. The efficacy of skip connections in this context remains to be fully explored. While transformer-based approaches offer another avenue, their substantial data requirements may render them less practical for the current application. The third potential strategy involves feature pooling followed by neural network training, though this method risks information loss as well. It is worth noting that the QA component could potentially inform the ensemble about its confidence in each input. This approach might allow for more effective utilization of high-quality inputs while naturally reducing the influence of poor-quality segmentations. Furthermore, it may be valuable to investigate whether the ensemble can inherently assign appropriate weights to inputs of varying quality, potentially obviating the need for explicit filtering.

Given these considerations, the first approach — utilizing a maximum number of inputs with an ignore mask and weight sharing — appears to be the most promising direction for initial exploration. This method offers flexibility in handling variable inputs while maintaining a consistent model structure. However, a comprehensive examination of all proposed approaches would provide a robust foundation for the final methodological decision. For the initial implementation, inputs of fixed size could be utilized, with the potential for exploring variable-sized inputs in future iterations of the model. This proposed architecture can be seen in Figure 4.3.

4.5 New Integrated QA-Ensemble Pipeline Architecture

The proposed approach integrates QA and dynamic ensemble methods into a multi-stage pipeline for image segmentation. As multiple approaches are explored, the pipeline will change and adapt, but a general overview can be seen in Figure 4.4.

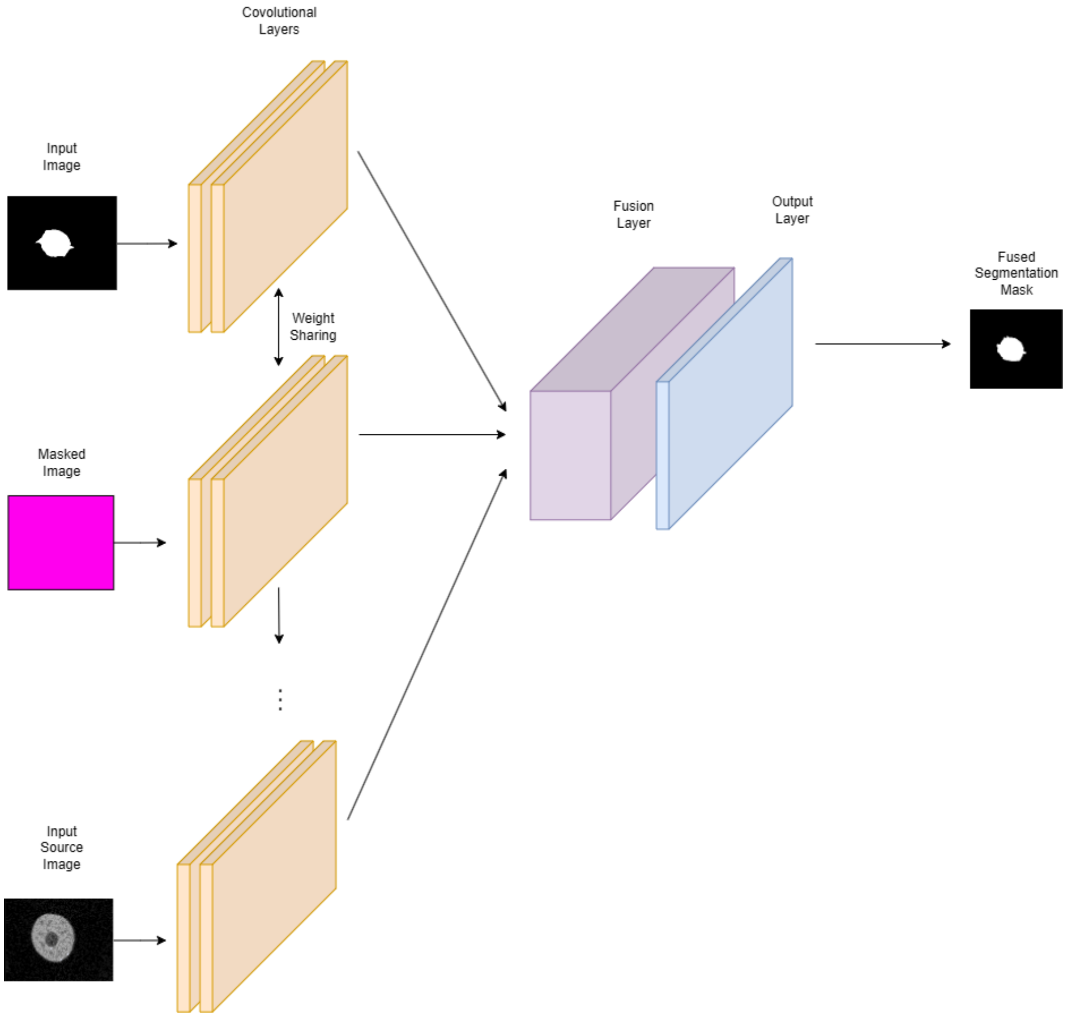


Figure 4.3: Scheme of the proposed dynamic ensemble utilizing Convolutional Neural Networks with weight sharing and excess input masking.

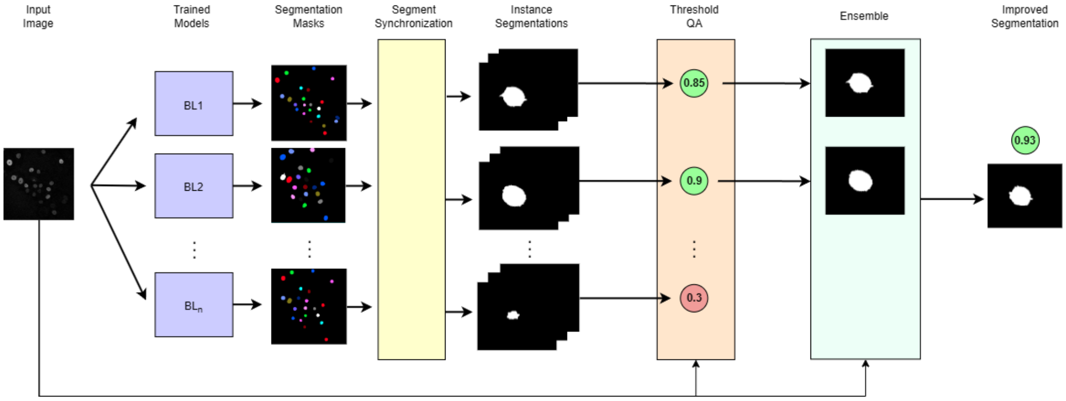


Figure 4.4: Scheme of the proposed dynamic ensemble of segmentation models with Quality Assurance filter. Base Learners (BL) segment the image, and segmentation instances are then synchronized between models. A QA method evaluates these segmentations and passes them to an ensemble method that produces improved segmentation.

Chapter 5

Conclusion and Future Work

This research introduces a novel approach to enhancing machine learning model performance through the integration of QA and dynamic ensemble methods. A sophisticated QA network has been proposed that considers both frame-by-frame and temporal consistency, alongside a dynamic ensemble architecture capable of handling variable inputs. These components have been combined into a unified pipeline that promises to improve results beyond current state-of-the-art methods across various domains.

The architecture's performance will first be assessed using the cell tracking challenge datasets. This evaluation will involve comparing the proposed method against state-of-the-art individual models and static ensemble methods. The comparison will utilize standard metrics including Intersection over Union (IoU), Dice Coefficient, and F1 score. To ensure unbiased evaluation, the assessment will be conducted on a separate test set not seen during the training phase. Furthermore, each component of the architecture will be tested individually to quantify its specific impact on the overall performance. This comprehensive evaluation approach will provide a clear understanding of the proposed method's effectiveness relative to existing techniques. It will also offer insights into which components contribute most significantly to any observed improvements, guiding future refinements and applications of the method.

Insights gained from this research may help further improve trust in machine learning approaches and provide more accurate models. The improved segmentations resulting from the work on cell tracking challenge could be utilized as silver truth data for training new models. This approach would allow for the creation of larger, more diverse datasets, potentially leading to more robust and generalizable segmentation models. The increased quantity and quality of training data could enhance the performance of deep learning models.

This research lays the groundwork for several promising avenues of future investigation:

Generalization to Other Computer Vision Domains While our focus has been on cell tracking, future work could explore the applicability of our methods to other image segmentation tasks, such as medical imaging or autonomous driving.

Application to Dynamic Simulation Reweighting The QA and dynamic ensemble techniques developed in this work could be adapted for machine learning based approaches to problems beyond computer vision, such as reweighting of dynamic simulations. This approach could potentially improve the accuracy and efficiency of simulations in fields such as molecular dynamics, fluid dynamics, or climate modeling.

These directions aim to extend the impact of our work beyond its current scope, potentially influencing a broader range of scientific and practical applications.

Bibliography

1. HARALICK, Robert M; SHAPIRO, Linda G. Image segmentation techniques. *Computer vision, graphics, and image processing.* 1985, vol. 29, no. 1, pp. 100–132.
2. JUNG, Sunguk; HEO, Hyeonbeom; PARK, Sangheon; JUNG, Sung-Uk; LEE, Kyungjae. Benchmarking deep learning models for instance segmentation. *Applied Sciences.* 2022, vol. 12, no. 17, p. 8856.
3. MINAEE, Shervin; BOYKOV, Yuri; PORIKLI, Fatih; PLAZA, Antonio; KEHTARNAVAZ, Nasser; TERZOPOULOS, Demetri. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence.* 2021, vol. 44, no. 7, pp. 3523–3542.
4. LONG, Jonathan; SHELHAMER, Evan; DARRELL, Trevor. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2015, pp. 3431–3440.
5. HE, Kaiming; GKIOXARI, Georgia; DOLLÁR, Piotr; GIRSHICK, Ross. Mask r-cnn. In: *Proceedings of the IEEE international conference on computer vision.* 2017, pp. 2961–2969.
6. CHEN, Liang-Chieh; PAPANDREOU, George; KOKKINOS, Iasonas; MURPHY, Kevin; YUILLE, Alan L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence.* 2017, vol. 40, no. 4, pp. 834–848.
7. RONNEBERGER, Olaf; FISCHER, Philipp; BROX, Thomas. U-net: Convolutional networks for biomedical image segmentation. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18.* Springer, 2015, pp. 234–241.
8. TERVEN, Juan; CÓRDOVA-ESPARZA, Diana-Margarita; ROMERO-GONZÁLEZ, Julio-Alejandro. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction.* 2023, vol. 5, no. 4, pp. 1680–1716.

9. KIRILLOV, Alexander; MINTUN, Eric; RAVI, Nikhila; MAO, Hanzi; ROLLAND, Chloe; GUSTAFSON, Laura; XIAO, Tete; WHITEHEAD, Spencer; BERG, Alexander C; LO, Wan-Yen, et al. Segment anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026.

10. SCHMIDT, Uwe; WEIGERT, Martin; BROADDUS, Coleman; MYERS, Gene. Cell detection with star-convex polygons. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*. Springer, 2018, pp. 265–273.

11. FENG, Di; HAASE-SCHÜTZ, Christian; ROSENBAUM, Lars; HERTLEIN, Heinz; GLAESER, Claudius; TIMM, Fabian; WIESBECK, Werner; DIETMAYER, Klaus. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*. 2020, vol. 22, no. 3, pp. 1341–1360.

12. MOHANTY, Sharada Prasanna; CZAKON, Jakub; KACZMAREK, Kamil A; PYSKIR, Andrzej; TARASIEWICZ, Piotr; KUNWAR, Saket; ROHRBACH, Janick; LUO, Dave; PRASAD, Manjunath; FLEER, Sascha, et al. Deep learning for understanding satellite imagery: An experimental survey. *Frontiers in Artificial Intelligence*. 2020, vol. 3, p. 534696.

13. LAMBERT, Benjamin; FORBES, Florence; DOYLE, Senan; DEHAENE, Harmonie; DOJAT, Michel. Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. *Artificial Intelligence in Medicine*. 2024, p. 102830.

14. KAISER, Timo; ULMAN, Vladimir; ROSENHAHN, Bodo. CHOTA: A Higher Order Accuracy Metric for Cell Tracking. *arXiv preprint arXiv:2408.11571*. 2024.

15. CHABRIER, Sébastien; EMILE, Bruno; LAURENT, Hèlène; ROSENBERGER, Christophe; MARCHE, P. Unsupervised evaluation of image segmentation application to multi-spectral images. In: *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004*. IEEE, 2004, vol. 1, pp. 576–579.

16. AREGA, Tewodros Weldebirhan; BRICQ, Stéphanie; LEGRAND, François; JACQUIER, Alexis; LALANDE, Alain; MERIAUDEAU, Fabrice. Automatic uncertainty-based quality controlled T1 mapping and ECV analysis from native and post-contrast cardiac T1 mapping images using Bayesian vision transformer. *Medical image analysis*. 2023, vol. 86, p. 102773.

17. DEVRIES, Terrance; TAYLOR, Graham W. Leveraging uncertainty estimates for predicting segmentation quality. *arXiv preprint arXiv:1807.00502*. 2018.

18. CHEN, Xinyuan; MEN, Kuo; CHEN, Bo; TANG, Yu; ZHANG, Tao; WANG, Shulian; LI, Yexiong; DAI, Jianrong. CNN-based quality assurance for automatic segmentation of breast cancer in radiotherapy. *Frontiers in Oncology*. 2020, vol. 10, p. 524.

19. LIN, Qiao; CHEN, Xin; CHEN, Chao; GARIBALDI, Jonathan M. A novel quality control algorithm for medical image segmentation based on fuzzy uncertainty. *IEEE Transactions on Fuzzy Systems.* 2022, vol. 31, no. 8, pp. 2532–2544.

20. ARBELLE, Assaf; ELUL, Eliav; RAVIV, Tammy Riklin. QANet–Quality Assurance Network for Image Segmentation. *arXiv preprint arXiv:1904.08503.* 2019.

21. ZAMAN, Fahim Ahmed; ZHANG, Lichun; ZHANG, Honghai; SONKA, Milan; WU, Xiaodong. Segmentation quality assessment by automated detection of erroneous surface regions in medical images. *Computers in biology and medicine.* 2023, vol. 164, p. 107324.

22. YAZDAN-PANAH, Arya; STANKOFF, Bruno; COLLIOT, Olivier. Automatic quality control of segmentation results using early epochs as data augmentation: application to choroid plexuses. In: *Medical Imaging 2024: Image Processing.* SPIE, 2024, vol. 12926, pp. 335–340.

23. ZHOU, Leixin; DENG, Wenxiang; WU, Xiaodong. *Robust Image Segmentation Quality Assessment.* 2020. Available from arXiv: `1903.08773 [cs.CV]`.

24. ZAMAN, Fahim Ahmed; WU, Xiaodong; XU, Weiyu; SONKA, Milan; MUDUMBAI, Raghuraman. Trust, but Verify: Robust Image Segmentation using Deep Learning. In: *2023 57th Asilomar Conference on Signals, Systems, and Computers.* IEEE, 2023, pp. 1070–1074.

25. LI, Kang; YU, Lequan; HENG, Pheng-Ann. Towards reliable cardiac image segmentation: Assessing image-level and pixel-level segmentation quality via self-reflective references. *Medical Image Analysis.* 2022, vol. 78, p. 102426.

26. ZHANG, Yizhe; WANG, Shuo; ZHOU, Tao; DOU, Qi; CHEN, Danny Z. SQA-SAM: Segmentation Quality Assessment for Medical Images Utilizing the Segment Anything Model. *arXiv preprint arXiv:2312.09899.* 2023.

27. VALINDRIA, Vanya V; LAVDAS, Ioannis; BAI, Wenjia; KAMNITSAS, Konstantinos; ABOAGYE, Eric O; ROCKALL, Andrea G; RUECKERT, Daniel; GLOCKER, Ben. Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. *IEEE transactions on medical imaging.* 2017, vol. 36, no. 8, pp. 1597–1606.

28. ROBINSON, Robert; OKTAY, Ozan; BAI, Wenjia; VALINDRIA, Vanya V; SANGHVI, Mihir M; AUNG, Nay; PAIVA, José M; ZEMRAK, Filip; FUNG, Kenneth; LUKASCHUK, Elena, et al. Real-time prediction of segmentation quality. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11.* Springer, 2018, pp. 578–585.

29. ALTMAN, MB; KAVANAUGH, JA; WOOTEN, HO; GREEN, OL; DEWEES, TA; GAY, H; THORSTAD, WL; LI, H; MUTIC, S. A framework for automated contour quality assurance in radiation therapy including adaptive techniques. *Physics in Medicine & Biology.* 2015, vol. 60, no. 13, p. 5199.

30. CHEN, Hsin-Chen; TAN, Jun; DOLLY, Steven; KAVANAUGH, James; ANASTASIO, Mark A; LOW, Daniel A; HAROLD LI, H; ALTMAN, Michael; GAY, Hiram; THORSTAD, Wade L, et al. Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: a general strategy. *Medical physics*. 2015, vol. 42, no. 2, pp. 1048–1059.

31. ZHANG, Ying; PLAUTZ, Tia E; HAO, Yao; KINCHEN, Catherine; LI, X Allen. Texture-based, automatic contour validation for online adaptive replanning: a feasibility study on abdominal organs. *Medical physics*. 2019, vol. 46, no. 9, pp. 4010–4020.

32. THAMBAWITA, Vajira; HICKS, Steven A; HALVORSEN, Pål; RIEGLER, Michael A. Divergentnets: Medical image segmentation by network ensemble. *arXiv preprint arXiv:2107.00283*. 2021.

33. JIA, Jianguo; LIANG, Wen; LIANG, Youzhi. A review of hybrid and ensemble in deep learning for natural language processing. *arXiv preprint arXiv:2312.05589*. 2023.

34. CAO, Yue; GEDDES, Thomas Andrew; YANG, Jean Yee Hwa; YANG, Pengyi. Ensemble deep learning in bioinformatics. *Nature Machine Intelligence*. 2020, vol. 2, no. 9, pp. 500–508.

35. FERNÁNDEZ-DELGADO, Manuel; CERNADAS, Eva; BARRO, Senén; AMORIM, Dinani. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*. 2014, vol. 15, no. 1, pp. 3133–3181.

36. KO, Albert HR; SABOURIN, Robert; BRITTO JR, Alceu Souza. From dynamic classifier selection to dynamic ensemble selection. *Pattern recognition*. 2008, vol. 41, no. 5, pp. 1718–1731.

37. KROGH, Anders; VEDELSBY, Jesper. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*. 1994, vol. 7.

38. NANNI, Loris; LUMINI, Alessandra; FANTOZZI, Carlo. Exploring the potential of ensembles of deep learning networks for image segmentation. *Information*. 2023, vol. 14, no. 12, p. 657.

39. KANG, Jaeyong; GWAK, Jeonghwan. Ensemble of instance segmentation models for polyp segmentation in colonoscopy images. *IEEE Access*. 2019, vol. 7, pp. 26440–26447.

40. LYKSBORG, Mark; PUONTI, Oula; AGN, Mikael; LARSEN, Rasmus. An ensemble of 2D convolutional neural networks for tumor segmentation. In: *Image Analysis: 19th Scandinavian Conference, SCIA 2015, Copenhagen, Denmark, June 15-17, 2015. Proceedings 19*. Springer, 2015, pp. 201–211.

41. CANALINI, Laura; POLLASTRI, Federico; BOLELLI, Federico; CANCILLA, Michele; ALLEGRETTI, Stefano; GRANA, Costantino. Skin lesion segmentation ensemble with diverse training strategies. In: *Computer Analysis of Images and Patterns: 18th International Conference, CAIP 2019, Salerno, Italy, September 3–5, 2019, Proceedings, Part I 18*. Springer, 2019, pp. 89–101.

42. DANG, Truong; NGUYEN, Tien Thanh; MORENO-GARCIA, Carlos Francisco; ELYAN, Eyad; MCCALL, John. Weighted ensemble of deep learning models based on comprehensive learning particle swarm optimization for medical image segmentation. In: *2021 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2021, pp. 744–751.

43. CATALANO, Nico; MARANELLI, Alessandro; CHIATTI, Agnese; MATTEUCCI, Matteo. More than the Sum of Its Parts: Ensembling Backbone Networks for Few-Shot Segmentation. *arXiv preprint arXiv:2402.06581*. 2024.

44. CERPA, Alonso; MEZA-LOVON, Graciela; FERNÁNDEZ, Manuel E Loaiza. Ensemble Learning to Perform Instance Segmentation over Synthetic Data. In: *International Symposium on Visual Computing*. Springer, 2021, pp. 313–324.

45. WU, Liming; CHEN, Alain; SALAMA, Paul; DUNN, Kenneth W; DELP, Edward J. An ensemble learning and slice fusion strategy for three-dimensional nuclei instance segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 1884–1894.

46. KHIRODKAR, Rawal; SMITH, Brandon; CHANDRA, Siddhartha; AGRAWAL, Amit; CRIMINISI, Antonio. Sequential Ensembling for Semantic Segmentation. *arXiv preprint arXiv:2210.05387*. 2022.

47. HOU, Wen-hui; WANG, Xiao-kang; ZHANG, Hong-yu; WANG, Jian-qiang; LI, Lin. A novel dynamic ensemble selection classifier for an imbalanced data set: An application for credit risk assessment. *Knowledge-Based Systems*. 2020, vol. 208, p. 106462.

48. JIAO, Botao; GUO, Yinan; GONG, Dunwei; CHEN, Qiuju. Dynamic Ensemble Selection for Imbalanced Data Streams With Concept Drift. *IEEE Transactions on Neural Networks and Learning Systems*. 2024, vol. 35, no. 1, pp. 1278–1291. Available from DOI: `10.1109/TNNLS.2022.3183120`.

49. SWAMINATHAN, Bhuvaneswari; PALANI, Saravanan; VAIRAVASUNDARAM, Subramaniyaswamy. Meta learning-based dynamic ensemble model for crop selection. *Applied Artificial Intelligence*. 2022, vol. 36, no. 1, p. 2145646.

50. MARKATOPOULOU, Fotini; TSOUMAKAS, Grigorios; VLAHAVAS, Ioannis. Dynamic ensemble pruning based on multi-label classification. *Neurocomputing*. 2015, vol. 150, pp. 501–512.

51. BIAN, Yijun; WANG, Yijun; YAO, Yaqiang; CHEN, Huanhuan. Ensemble pruning based on objection maximization with a general distributed framework. *IEEE transactions on neural networks and learning systems*. 2019, vol. 31, no. 9, pp. 3766–3774.

52. PÉREZ-GODOY, Marıa D; MOLINA, Marta; MARTINEZ, Francisco; ELIZONDO, David; CHARTE, Francisco; RIVERA, Antonio J. DESReg: Dynamic Ensemble Selection library for Regression tasks. *Neurocomputing*. 2024, vol. 580, p. 127487.

53. CRUZ, Rafael MO; HAFEMANN, Luiz G; SABOURIN, Robert; CAVALCANTI, George DC. DESlib: A Dynamic ensemble selection library in Python. *Journal of Machine Learning Research*. 2020, vol. 21, no. 8, pp. 1–5.

54. HANN, Evan; POPESCU, Iulia A; ZHANG, Qiang; GONZALES, Ricardo A; BARUTÇU, Ahmet; NEUBAUER, Stefan; FERREIRA, Vanessa M; PIECHNIK, Stefan K. Deep neural network ensemble for on-the-fly quality control-driven segmentation of cardiac MRI T1 mapping. *Medical image analysis*. 2021, vol. 71, p. 102029.

55. HANN, Evan; GONZALES, Ricardo A; POPESCU, Iulia A; ZHANG, Qiang; FERREIRA, Vanessa M; PIECHNIK, Stefan K. Ensemble of deep convolutional neural networks with monte carlo dropout sampling for automated image segmentation quality control and robust deep learning using small datasets. In: *Annual Conference on Medical Image Understanding and Analysis*. Springer, 2021, pp. 280–293.

56. GARCIA, Salvador; ZHANG, Zhong-Liang; ALTALHI, Abdulrahman; ALSHOMRANI, Saleh; HERRERA, Francisco. Dynamic ensemble selection for multi-class imbalanced datasets. *Information Sciences*. 2018, vol. 445, pp. 22–37.

57. ZULFIRA, Fakhira Zahra; SUYANTO, Suyanto; SEPTIARINI, Anindita. Segmentation technique and dynamic ensemble selection to enhance glaucoma severity detection. *Computers in Biology and Medicine*. 2021, vol. 139, p. 104951.

58. GONZALES, Ricardo A; IBÁÑEZ, Daniel H; HANN, Evan; POPESCU, Iulia A; BURRAGE, Matthew K; LEE, Yung P; ALTUN, İbrahim; WEINTRAUB, William S; KWONG, Raymond Y; KRAMER, Christopher M, et al. Quality control-driven deep ensemble for accountable automated segmentation of cardiac magnetic resonance LGE and VNE images. *Frontiers in Cardiovascular Medicine*. 2023, vol. 10, p. 1213290.

59. LAD, Vedang; MUELLER, Jonas. Estimating label quality and errors in semantic segmentation data via any model. *arXiv preprint arXiv:2307.05080*. 2023.

60. MAŠKA, Martin; ULMAN, Vladimır; DELGADO-RODRIGUEZ, Pablo; GÓMEZ-DE-MARISCAL, Estibaliz; NEČASOVÁ, Tereza; GUERRERO PEÑA, Fidel A; REN, Tsang Ing; MEYEROWITZ, Elliot M; SCHERR, Tim; LÖFFLER, Katharina, et al. The cell tracking challenge: 10 years of objective benchmarking. *Nature Methods*. 2023, vol. 20, no. 7, pp. 1010–1020.

61. ABRAMOVA, Vera; LEAL ALVARADO, Vanessa; HILL, Martin; SMEJKALOVA, Tereza; MALY, Michal; VALES, Karel; DITTERT, Ivan; BOZIKOVA, Paulina; KYSILOV, Bohdan; HRCKA KRAUSOVA, Barbora, et al. Effects of Pregnanolone Glutamate and Its Metabolites on GABAA and NMDA Receptors and Zebrafish Behavior. *ACS Chemical Neuroscience*. 2023, vol. 14, no. 10, pp. 1870–1883.

62. HURTIK, Petr; ČIŽ, David; KALÁB, Oto; MUSIOLEK, David; KOČÁREK, Petr; TOMIS, Martin. Software for visual insect tracking based on F-transform pattern matching. In: *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. IEEE, 2018, pp. 528–533.

63. CHICCO, Davide. Siamese neural networks: An overview. *Artificial neural networks*. 2021, pp. 73–94.

64. ARBELLE, Assaf; RAVIV, Tammy Riklin. Microscopy cell segmentation via convolutional LSTM networks. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1008–1012.